



Diversité génomique des bactéries pathogènes du complexe d'espèces *Borrelia burgdorferi*: évolution et épidémiologie moléculaire

Maude Pithon Jacquot

► To cite this version:

Maude Pithon Jacquot. Diversité génomique des bactéries pathogènes du complexe d'espèces *Borrelia burgdorferi*: évolution et épidémiologie moléculaire. Sciences agricoles. Université Blaise Pascal - Clermont-Ferrand II, 2014. Français. NNT : 2014CLF22494 . tel-01164371

HAL Id: tel-01164371

<https://theses.hal.science/tel-01164371>

Submitted on 16 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**ECOLE DOCTORALE SCIENCES DE LA VIE, SANTE, AGRONOMIE,
ENVIRONNEMENT**
N° d'ordre : 640

THESE
Présentée à l'INRA de Clermont-Ferrand – Theix
pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ BLAISE PASCAL DE CLERMONT-FERRAND

Spécialité : Physiologie et génétique moléculaires (Biologie évolutive et épidémiologie)

**Diversité génomique des bactéries pathogènes
du complexe d'espèces *Borrelia burgdorferi* :
évolution et épidémiologie moléculaire**

par
Maude JACQUOT (PITHON)

Directeur de thèse : **Christian Ducrot**
Encadrant : **Xavier Bailly**

Soutenue publiquement le 8 octobre 2014 devant le jury composé de :

Roman Biek, Senior Lecturer, Université de Glasgow, Glasgow (Écosse)	Rapporteur
Thierry Wirth, Professeur, EPHE/MNHN, Paris	Rapporteur
Karen McCoy, Directrice de recherche, CNRS/IRD, Montpellier	Examinatrice
Frédéric Delbac, Professeur, Université Blaise Pascal, Clermont-Ferrand	Examineur
Virginie Ravigné, Chargé de recherche, Cirad, Montpellier	Examinatrice
Christian Ducrot, Directeur de recherche, INRA, Clermont-Ferrand - Theix	Directeur de thèse
Xavier Bailly, Ingénieur de recherche, INRA, Clermont-Ferrand – Theix	Encadrant de thèse

A mon papi Michel

Remerciements

Je souhaite remercier du fond du cœur toutes les personnes ayant participé de près ou de loin à l'aventure scientifique et humaine que représente ces trois ans de thèse.

Je commencerai par le plus important. Xavier, merci pour ces trois, presque quatre, années passées avec à vos côtés. C'était un vrai bonheur de travailler avec vous. Merci pour toute la joie, la bonne humeur, l'enthousiasme et aussi la pédagogie que vous mettez dans tout ce que vous entreprenez. On nous a souvent, paraît-il, entendu rire jusque dans le couloir : croyez-moi, je n'oublierai donc pas que si l'on fait tout ça, c'est juste « Pour le plaisir... ». Merci de m'avoir accordé votre confiance dès le départ de cette thèse malgré le peu de connaissances que je pouvais avoir en évolution, en génétique des populations et en épidémiologie. Merci pour tout ce que vous m'avez apporté, je repars plus confiante en moi et avec le sentiment d'avoir appris des milliers de choses ! Je ferai de mon mieux pour faire bon usage de tout cela dans le futur. De tout mon cœur, j'espère que bien d'autres que moi auront le privilège de vous avoir pour encadrant voire même comme directeur de thèse (si vous m'invitez, je viendrai peut-être, quand même, vous voir quand vous passerez votre HDR:)...). En tout cas, une chose est sûre, où que le vent m'emmène, mon boss va très souvent me manquer.

Christian, merci de nous avoir accordé à Xavier et à moi toute votre confiance, tout en gardant toujours, un œil bienveillant à mon égard. Merci pour votre écoute et votre intérêt malgré la distance de mon sujet de thèse avec vos domaines de recherche. Je continuerai à penser à vous chaque fois que je changerai de couleur ou de coupe de cheveux et que je me rendrai pour la première fois au travail avec ! Personne ne commentera jamais mes extravagances aussi bien que vous. Merci pour votre gentillesse et vos attentions.

Je remercie très chaleureusement Roman Biek, Thierry Wirth, Karen McCoy, Frédéric Delbac et Virginie Ravigné pour avoir accepté de faire partie de mon jury de thèse et d'évaluer mon travail.

Merci également aux membres de mes comités de thèse, Karen McCoy, Olivier Plantard et Olivier Tenaillon, pour leurs remarques et leurs conseils précieux qui ont permis de faire avancer la réflexion sur mon sujet.

Je remercie toutes les personnes avec qui j'ai eu l'occasion de collaborer, de près ou de loin. Merci à Elisabeth Ferquel et Valérie Choumet pour leur participation au projet de génomique et pour nous avoir reçu à plusieurs reprises à l'Institut Pasteur. Merci à Maud Marsot et Jean-Louis Chapuis pour leur collaboration dans le cadre du projet sur les rongeurs de Sénart.

Un grand merci également à chaque membre de l'unité d'Épidémiologie Animale. Vous avez tous quelque chose d'extraordinaire et je vous remercie d'avoir partagé avec moi l'atmosphère qui règne dans l'unité et qui fait qu'on s'y sent bien. Prenez bien soin de cette ambiance qui est si précieuse.

Tout particulièrement, je remercie Gwenaël et Christian, duo de choc à la tête de l'unité pendant ces trois ans de thèse. Merci de m'avoir accueillie parmi les membres de l'équipe, ce fut un plaisir de venir travailler chaque matin. Gwenaël, ton humanité et toutes les attentions que tu portes à chacun, sont des qualités que je regrette déjà.

Jocelyn et David, je vous remercie pour tout le temps que vous m'avez accordé et pour la patience dont vous avez fait preuve pour me rendre service. La grille et mes simulations de

coalescent n'ont pas toujours fait bon ménage mais grâce à vous, nous avons fini par obtenir de jolis résultats. David, un jour je saurai utiliser Q-GIS correctement, et je penserai bien à toi, avec qui j'aurai fait mes premiers pas.

Patrick, merci de ne pas nous avoir fermé la porte au nez quand, n'importe quand, Xavier et moi arrivions dans votre bureau avec tant de questions. Les statistiques paraissent tellement plus simples quand vous êtes dans les environs...

Merci à toi Myriam pour ta gentillesse et ta douceur. Merci aussi d'avoir relu l'article de génomique à plusieurs reprises et pour y avoir apporté de la clarté pour les non-spécialistes.

Sébastien, merci pour ta bonne humeur que tu sais partager. Merci aussi à toi Elsa pour tous les bons conseils que tu donnes toujours. Anne-Sophie, merci pour ton aide précieuse pour l'acquisition des articles pour lesquels il me manquait les codes « magiques » d'accès aux revues. Françoise, je te remercie très fort pour ta disponibilité, ta patience et ton efficacité pour gérer tous les problèmes administratifs que l'on peut rencontrer.

Valérie, Nelly, merci de tout mon cœur pour votre soutien. L'une comme l'autre avez toujours été à mon écoute dans les bons comme dans les moins bons moments et j'ai trouvé certains jours en vous, la maman qu'il nous manque parfois au travail. J'espère recevoir de vos nouvelles souvent.

Énorme merci à toi Séverine. Il y a des rencontres dans la vie qui nous marquent et que l'on n'oublie jamais. Tu feras partie de ces personnes pour moi. Merci de m'avoir fait partager ton expérience de l'unité quand j'en avais besoin, elle m'a bien souvent permis de prendre du recul. Mais, c'est pour ta bienveillance, ta douceur et ton amitié que je te remercie le plus fort. Je te souhaite tout le meilleur pour ta vie de demain.

Élise, Amélie, Angélique, et Sylvain... Quel bonheur et surtout quelle chance d'avoir pu partager mon bureau avec vous pendant tout ce temps. C'était vraiment un plaisir d'échanger chaque jour avec vous, qu'il s'agisse de mots, de pauses clopes, de chocolat ou de thé. Lorsque l'on commence une thèse, je ne crois pas que l'on puisse rêver mieux, pour ce qui est de l'ambiance dans laquelle on va évoluer, alors merci. J'espère que nos routes se croiseront encore très souvent pour le travail mais aussi et surtout pour le plaisir de se retrouver. Élise, je te souhaite de tout cœur de t'épanouir dans ta nouvelle vie et de toujours t'émerveiller de ces bêtes bizarres que tu adores. Amélie, ne change rien, même sans boucle d'oreille et malgré ce qu'on a pu en dire pour te charrier, tu es tip top comme ça. Angélique, s'il te plaît, garde ton sourire et ton caractère haut en couleur que j'adore. Sylvain, prends bien soin de notre bureau, et surtout fais partager cet humour qui te caractérise tant encore longtemps ; il est vraiment génial. Personne ne vous égalera et mes co-burottes vont, c'est sûr, désespérément me manquer quand il me faudra m'installer ailleurs, dans un nouveau bureau.

Les filles, j'ai adoré partager avec vous les soirées zumba ou apéro au H-Park. Bien que l'on ait toujours « essayé de faire quelque chose », je ne suis pas sûre d'avoir travaillé quelques muscles que ce soit, à l'exception de mes zygomatiques bien sûr ! Merci, c'était vraiment chouette.

Je ne serai jamais parvenue au bout de ces huit années d'études sans le soutien sans faille de mes amis les plus proches et de ma famille que j'aime fort.

Ludivine, Xavier, nous sommes parvenus au bout de cette aventure qu'est la thèse, même si tu nous as lâché, Xav ! Comme tout ce que nous faisons avec vous, c'était un bonheur de partager ces moments, alors merci. Merci de m'avoir écouté vous raconter mes histoires de labo alors que vous ne connaissiez personne, pour avoir compati quand j'en avais besoin ou relativisé lorsque cela s'avérait nécessaire. Une page se tourne et nous allons prendre chacun des chemins différents, mais je sais qu'il nous suffira de nous retourner pour vous savoir juste là, pas loin.

Papa, Maman, Auré, merci d'avoir toujours cru en moi sans jamais douter, merci pour m'avoir encouragé et merci d'avoir été là pour m'accompagner à chaque moment important de la vie. Merci à mon doudou Maxence, soleil d'une vie, pour toute l'énergie qu'il me transmet et qui m'est indispensable. Merci à toi Carole pour ta douceur apaisante en toute circonstance.

Merci à vous, ma belle maman, pour vos encouragements et pour la fierté que je lis dans vos yeux lorsque vous nous regardez, votre fils et moi. Je suis toujours très touchée de tout ce que vous nous apportez.

Je remercie aussi très fort mes mamies et mes papis. Quelle chance de vous avoir encore tous les quatre, merci de votre soutien et de votre amour inconditionnel. Papi Michel, tu es sans aucun doute celui à qui je suis la plus fière de dire que j'ai réussi et c'est à toi que je dédie cette thèse. Merci de toujours avoir été là et... même si pour moi tu es éternel, je t'en prie, reste encore un peu.

A toi mon mari, merci pour tout ton amour qui chaque jour m'a donné tout ce dont j'avais besoin pour avancer et me sentir bien. Rappelle toi notre première année de fac, nous avons passé plus de temps à se lancer des boulettes de papiers et à faire la fête qu'à travailler mais finalement, l'un à côté de l'autre, nous sommes arrivés là où nous le souhaitons. Ce long parcours n'aurait eu que très peu de sens sans toi. Merci d'avoir supporté mes humeurs, mes doutes, mes colères, mes peines et, mes joies et euphories tout ce temps. Merci de me donner aujourd'hui le plus cadeau du monde... Je te dois tellement...

Table des matières

Liste des figures	p. 7
Introduction	p. 9
1. Les maladies infectieuses à transmission vectorielle	p. 11
1.1. Importance des maladies infectieuses à transmission vectorielle	p. 11
1.2. Cycles épidémiologiques complexes des maladies à transmission vectorielle	p. 13
1.2.1. Premiers acteurs du cycle : les agents infectieux	p. 13
1.2.2. Seconds acteurs du cycle : les hôtes	p. 16
1.2.3. Troisièmes acteurs du cycle : les vecteurs	p. 19
2. La biologie moléculaire et la génétique des populations au service de l'étude des maladies vectorielles	p. 22
2.1. Intégration des concepts phylogénétiques dans les approches d'épidémiologie moléculaire	p. 23
2.2. Apport de la génétique des populations au développement des approches phylodynamiques	p. 26
3. Modèle biologique : la borréliose de Lyme	p. 31
3.1. Le pathogène	p. 32
3.2. Les hôtes	p. 34
3.3. Le vecteur tique	p. 35
3.4. Structure et dynamique des génomes	p. 37
4. Problématiques et plan de thèse	p. 39
Chapitre 1 : Génomique comparative du complexe d'espèces <i>B. burgdorferi</i> s.l	p. 43
Résumé de l'article 1	p. 45
Article 1	p. 47
Données supplémentaires de l'article 1	p. 69
Chapitre 2 : Différenciation hôte-spécifique entre les populations de <i>B. burgdorferi</i> s.l.	p. 81
Résumé de l'article 2	p. 83
Article 2	p. 85
Données supplémentaires de l'article 2	p. 97

Chapitre 3 : Cycles de transmission des génotypes de <i>B. burgdorferi</i> s.l. et assignation des espèces d'hôtes à ces cycles	p. 105
Résumé de l'article 3	p. 107
Article 3	p. 109
Données supplémentaires de l'article 3	p. 141
Discussion et perspectives	p. 143
1. Isolement génétique des différentes génoespèces de <i>B. burgdorferi</i> s.l. et recombinaison	p. 145
2. Origine de l'émergence de nouveaux groupes de <i>Borrelia</i> et adaptation à un nouvel environnement	p. 148
3. Circulation des populations de <i>B. burgdorferi</i> s.l. entre les hôtes et les vecteurs	p. 149
3.1. Dynamiques de circulation hôtes spécifiques des différentes génoespèces de <i>Borrelia</i>	p. 149
3.2. Contribution des différentes espèces d'hôtes	p. 153
4. Conclusion	p. 157
Références bibliographiques	p. 159

Liste des figures

Figure 1 : Distribution spatiale des cas reportés de dengue dans le monde par décade depuis 1943 [Messina *et al.*, 2014]

Figure 2 : Les différents facteurs qui ont le potentiel d'influencer le statut des maladies à transmission vectorielle [Sutherst, 2004]

Figure 3 : Taux de mutation par site nucléotidique en fonction de la taille des génomes pour différentes entités biologiques [Gago *et al.*, 2009]

Figure 4 : Les systèmes cible-réservoir [Haydon *et al.*, 2002]

Figure 5 : Schématisation d'un modèle SIR

Figure 6 : La circulation des agents pathogènes au sein d'une communauté d'hôte est dépendante de la spécialisation des vecteurs [McCoy *et al.*, 2013]

Figure 7 : La résolution est dépendante du type et du nombres de marqueurs utilisés [Maiden *et al.*, 2013]

Figure 8 : Principe de l'approche Bayésienne et inférences

Figure 9 : Illustration de différents scénarii évolutifs pouvant être testé via une analyse de coalescent [Grünwald and Goss, 2011]

Figure 10 : Evolution de la valeur du D de Tajima au cours des générations après un goulot d'étranglement [Fay and Wu, 1999]

Figure 11 : Différenciation génétique, exemple de comportement de l'indice de fixation F_{ST}

Figure 12 : Déséquilibre de liaison et recombinaison [Rafalski, 2002]

Figure 13 : Structure épidémique [Maynard Smith *et al.*, 2000]

Figure 14 : Spirochètes du genre *Borrelia* [Brisson *et al.*, 2012]

Figure 15 : Distribution géographique de *B. burgdorferi* s.l. aux Etats-Unis et en Europe [Franke *et al.*, 2013]

Figure 16 : Distribution géographique de *I. ricinus* s.l. [Stanek *et al.*, 2012]

Figure 17 : Cycle de vie des tiques *I. ricinus* associé à celui de *B. burgdorferi* s.l. [Brisson *et al.*, 2012]

Figure 18 : Schématisation d'un modèle phylodynamique qui prendrait en compte la spécialisation adaptative des *Borrelia*.

Figure 19 : Distribution du nombre de génotypes détectés chez les individus hôtes aux deux loci *rplB* et *ospC*

Figure 20 : Co-occurrence des génotypes identifiés chez les tiques et les hôtes et des identifications de repas sanguin des tiques

INTRODUCTION

Les maladies infectieuses sont une des causes les plus importantes de morbidité chez l'homme et l'animal avec des conséquences à la fois économiques, sanitaires et écologiques. L'étude de la diversité des pathogènes qui les causent et de leurs dynamiques de circulation au sein des communautés d'hôtes et de vecteurs, peut fournir des informations importantes pour la prévention et le contrôle de ces maladies. Au cours de ma thèse, je me suis intéressée à la diversité génétique ainsi qu'aux cycles de transmission de l'agent pathogène responsable de la maladie de Lyme. Cette maladie est causée par les bactéries du complexe d'espèces *Borrelia burgdorferi* sensu lato (s.l.) transmises par les tiques lors de repas sanguin. Elles sont capables d'infecter plusieurs espèces d'hôtes vertébrés, dont en particulier l'homme chez qui les symptômes peuvent être graves. Au cours de cette introduction, je présenterai, dans un premier temps, les propriétés des maladies infectieuses à transmission vectorielle qui posent questions et quels sont les enjeux actuels auxquels la recherche doit faire face pour mieux les surveiller et les contrôler. Pour cela, je décrirai tout d'abord quels sont les acteurs des cycles épidémiologiques complexes de ces maladies, puis les méthodes d'épidémiologie moléculaire dont on dispose à ce jour pour les étudier et les outils qui pourraient être d'avantage développés pour offrir de nouvelles opportunités. Enfin, je donnerai une description du modèle d'étude *B. burgdorferi* s.l. utilisé pour mes travaux et présenterai les problématiques et les questions abordées pendant ma thèse.

1. Les maladies infectieuses à transmission vectorielle

1.1. Importance des maladies infectieuses à transmission vectorielle

Les maladies infectieuses peuvent être transmises selon de nombreuses modalités telles que de la mère à l'enfant (transmission verticale), par contacts directs ou *via* l'environnement (transmission horizontale). Cependant, le mode de transmission *via* l'environnement faisant intervenir un vecteur est sans doute le plus complexe. Parmi les agents infectieux responsables des **maladies à transmission vectorielle**, il existe des protozoaires comme *Plasmodium* spp. qui causent le paludisme, des virus comme les *Flaviviridae* responsables de la dengue, ou des bactéries comme les *Borrelia* spp. responsables notamment de la borréliose de Lyme.

L'impact en **santé animale** des maladies à transmission vectorielle n'est pas des moindres. Les trypanosomiasés animales ou la fièvre catarrhale ovine sont des exemples de maladies à transmission vectorielle sans danger pour l'homme mais qui peuvent cependant

Introduction

causer d'importants dommages sur les élevages et induire des pertes économiques conséquentes. Par exemple, le coût net de la fièvre catarrhale ovine (sérotypage 8) aux Pays-Bas a été estimé à 32,4 millions d'euros pour l'épidémie de 2006 et entre 164 et 175 millions d'euros pour l'épidémie de 2007 [Velthuis *et al.*, 2010]. Plus important encore, les maladies à transmission vectorielle ont un fort impact en **santé humaine**. Plus d'un milliard de personnes dans le monde sont infectées chaque année par des maladies à transmission vectorielle et plus d'un million en meurent. Parmi ces maladies, celle qui fait le plus de victimes est le paludisme, maladie infectieuse à laquelle on attribue 8 % des décès chez les enfants de moins de cinq ans [Bryce *et al.*, 2005]. Toutefois la maladie à transmission vectorielle qui s'étend le plus rapidement dans le monde est la dengue dont l'incidence a été multipliée par 30 au cours des 50 dernières années [WHO, 2009] (Figure 1).

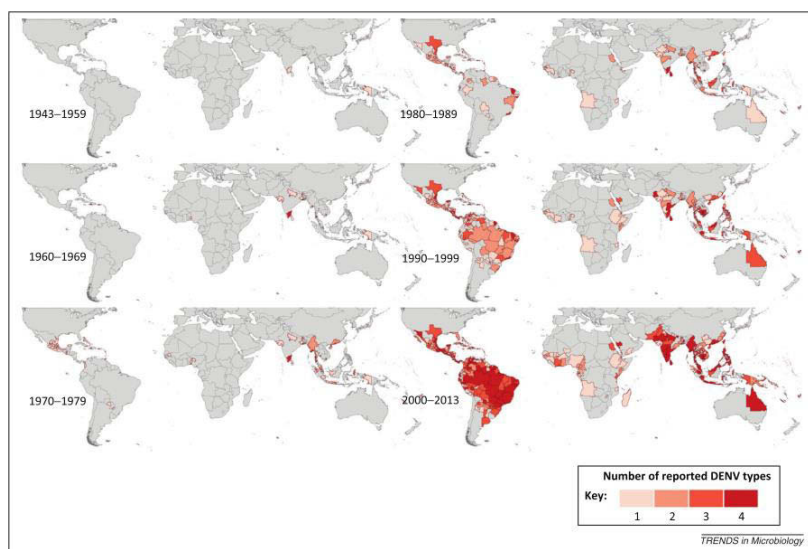


Figure 1 : Distribution spatiale des cas reportés de dengue dans le monde par décennie depuis 1943 [Messina *et al.*, 2014]

DENV types : Sérotypes du virus de la dengue

De nombreux facteurs semblent faciliter l'**expansion** de ces maladies [Sutherst, 2004] (Figure 2). Tout d'abord, l'intensification et la mondialisation des échanges de biens ainsi que les mouvements de personnes, fréquents et sur de longues distances, autorisent aujourd'hui le déplacement rapide des différents intervenants des cycles de transmission de ces maladies tels que les hôtes et les vecteurs infectieux. Ensuite, la modification des écosystèmes et de l'environnement liée d'une part aux activités humaines et d'autre part aux facteurs climatiques joue, et jouera dans les décennies à venir, un rôle important dans la dynamique et la diffusion des infections vectorielles. De fait, les maladies à transmission vectorielle représentent déjà

plus de 25 % des maladies infectieuses qualifiées d'émergentes [Jones *et al.*, 2008]. Les maladies infectieuses émergentes sont celles dont l'incidence est en augmentation suite à sa première introduction dans une nouvelle population d'hôtes. On qualifie plutôt de ré-émergentes, celles dont l'incidence augmente dans une population d'hôtes, à la suite de changements à long terme de leur épidémiologie sous-jacente [Woolhouse and Dye 2001 ; Woolhouse, 2002]. A une échelle géographique donnée, l'incidence d'une maladie, initialement émergente ou ré-émergente, peut atteindre un équilibre. Dans ce cas, l'épidémie s'est transformée en une endémie. Les endémies sont plus simples à surveiller et à étudier que les épidémies en raison de la stabilité du système qui s'avère alors plus facile à modéliser.

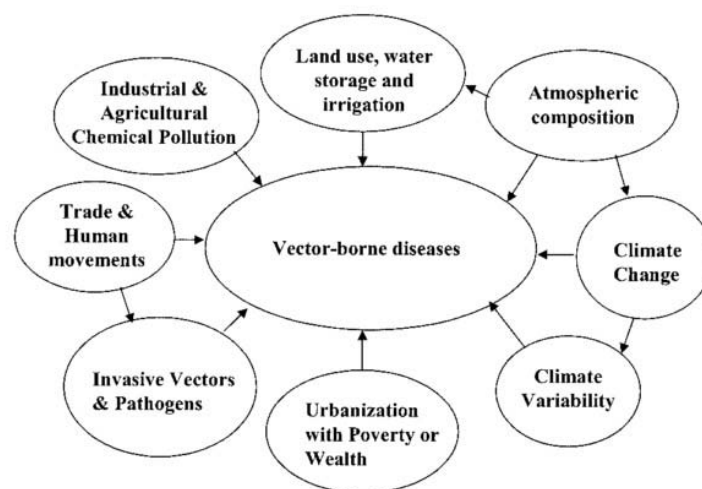


Figure 2 : Les différents facteurs qui ont le potentiel d'influencer le statut des maladies à transmission vectorielle [Sutherst, 2004]

1.2. Cycles épidémiologiques complexes des maladies à transmission vectorielle

La maîtrise des systèmes de maladie à transmission vectorielle constitue donc un enjeu majeur de santé publique animale et humaine. Elle passe notamment par la caractérisation des agents infectieux et la compréhension des mécanismes et paramètres de la transmission de l'infection qui peuvent être complexes, puisque le mode de transmission indirecte fait intervenir de nombreux acteurs. Les **agents** infectieux peuvent circuler chez plusieurs types d'**hôtes** et de **vecteurs** qui peuvent être impliqués à différentes échelles et niveaux du cycle épidémiologique des maladies.

1.2.1. Premiers acteurs du cycle : les agents infectieux

Un agent pathogène responsable d'une maladie est transmis aux hôtes en fonction d'un « **taux de transmission** » qui décrit le taux par unité de temps auquel les hôtes sensibles (non

Introduction

infectés et non immunisés) sont convertis en hôtes infectés par contact avec du matériel infectieux. Les taux de transmission font l'objet de nombreuses études car ils peuvent être hétérogènes en fonction des espèces hôtes considérées. Cependant, ils sont notoirement difficiles à quantifier. Les termes à prendre en compte pour leurs calculs font débat et différentes formulations mathématiques de la fonction de transmission existent [McCallum *et al.*, 2001 ; Begon *et al.*, 2002]. De façon la plus simple, le nombre de nouveaux infectés par unité de temps, dépend du nombre d'individus sensibles et de ce que l'on appelle communément la **force d'infection** notée λ . Elle est le produit de trois termes : i) du taux de contact pour un individu sensible c , ii) de la probabilité que ces contacts se fassent avec un individu infectieux (prévalence de l'infection) p , iii) de la probabilité de succès de la transmission ν [Begon *et al.*, 2002]. Le paramètre ν est considéré comme constant et p est équivalent à la proportion d'individus infectés sur la population totale de taille N . Ainsi, le paramètre le plus sujet à être modifié est le taux de contact qui peut dépendre à la fois de la proportion d'hôtes infectés dans la population (transmission fréquence-dépendante) et du nombre d'individus hôtes infectés (transmission densité-dépendante).

En épidémiologie, les agents infectieux sont également caractérisés par deux autres propriétés de grand intérêt : leur pathogénicité et leur virulence. Un agent est dit **pathogène** s'il provoque une maladie chez un organisme hôte c'est à dire l'apparition de manifestations cliniques, de symptômes. La **virulence** mesure, quant à elle, l'importance de ces symptômes en prenant en compte la rapidité avec laquelle se multiplie l'agent pathogène (pouvoir invasif) et sa capacité à sécréter des toxines (pouvoir toxique) [Calow, 1998]. Pathogénicité et virulence sont des paramètres épidémiologiques qui peuvent être variables d'un hôte à l'autre [Poulin and Combes, 1999] en fonction des espèces ou en réponse à des variations de susceptibilité des individus et/ou des populations concerné(e)s par l'infection. Ces paramètres sont également dépendants de l'environnement intra-hôte et particulièrement des événements de co-infections des individus par différents agents pathogènes. En premier lieu, les co-infections par des pathogènes de la même espèce ou d'espèces différentes peuvent conduire à une réduction des ressources disponibles pour l'hôte au profit des agents pathogènes et donc à une augmentation de la sévérité des symptômes [Frank, 1992 ; Frank, 1996 ; Nowak and May, 1994]. Ensuite, en raison de la compétition directe ou indirecte qui peut s'installer entre les pathogènes co-infectant un hôte, certaines études montrent que les souches les plus virulentes peuvent être favorisées [Cooper and Heinemann, 2005 ; de Roode *et al.* 2005] tandis que dans

d'autres cas, ce sont les souches les moins virulentes qui peuvent être favorisées, grâce à la mise en place de comportements coopératifs à leur profit, par exemple [Harrison *et al.*, 2006 ; Mayxay, 2004].

Taux de transmission, pathogénicité, et virulence sont donc les déterminants majeurs de la dynamique des maladies infectieuses et dépendent de la capacité des agents infectieux à s'adapter à leur environnement et aux modifications que cet environnement peut subir dans le cadre de co-infections ou de modification du statut immunitaire des hôtes par exemple. Ces **adaptations** passent par la génération de **diversité** et la modification des fréquences alléliques au sein des populations d'agents pathogènes. Les agents infectieux sont en effet des entités qui évoluent vite, avec un temps de génération souvent bien plus court que celui de leurs hôtes et des taux de mutations importants [Gago *et al.*, 2009] (Figure 3). Ils peuvent donc s'adapter rapidement aux pressions de sélection exercées par les hôtes. Par exemple, l'évolution rapide des antigènes est une des stratégies mises en place par les agents infectieux pour échapper au système immunitaire des hôtes qu'ils infectent. Ces évolutions peuvent modifier des caractéristiques clés pouvant favoriser l'émergence de maladies telles que la résistance contre les traitements ou l'adaptation à de nouvelles espèces d'hôtes [Reluga *et al.*, 2007]. De façon générale, les niches écologiques dans lesquelles les agents évoluent, façonnent un paysage adaptatif des variants génétiques dans lequel, les forces évolutives sont un moyen de circuler [Wright, 1932]. Ainsi, un des enjeux de la biologie évolutive appliquée aux pathogènes est de comprendre les processus qui génèrent et maintiennent la diversité de ces agents infectieux. La connaissance de ces processus peut être importante dans la mesure où elle permet i) de mieux comprendre la distribution et la dynamique des pathogènes et donc d'améliorer la prévention et le contrôle des maladies et ii) d'orienter les études de mise au point des vaccins ou de mise en place des stratégies d'adaptation des traitements selon les variants de l'agent pathogène.

Introduction

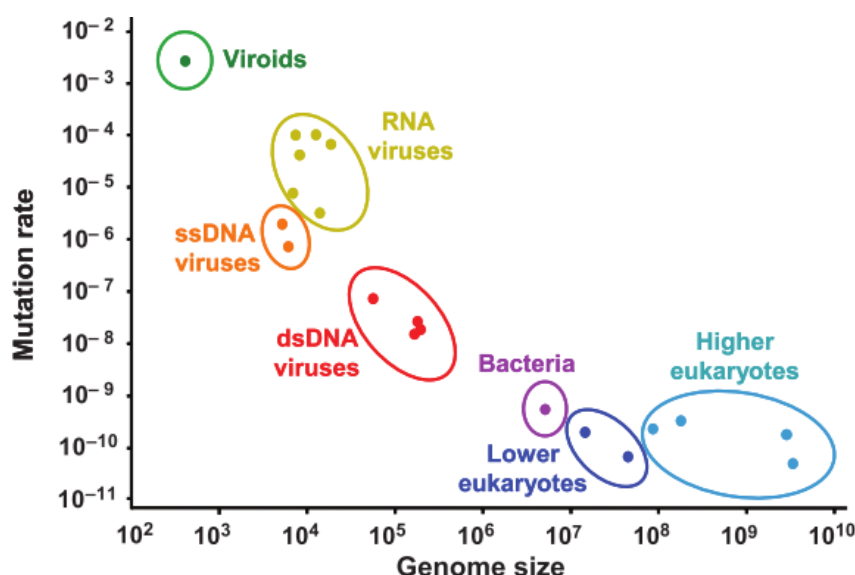


Figure 3 : Taux de mutation par site nucléotidique en fonction de la taille des génomes pour différentes entités biologiques [Gago *et al.*, 2009]

Pour les virus ARN, il s'agit du virus de la mosaïque du tabac, d'un rhinovirus humain, d'un poliovirus, du virus de la stomatite vésiculaire, du bactériophage $\Phi 6$, et du virus de la rougeole. Les virus à ADN simple brin sont les bactériophages $\Phi X174$ et m13. Les virus à ADN double brin sont le bactériophage λ , un virus de l'herpès simplex et les bactériophages T2 et T4. La bactérie est *Escherichia coli*. Les eucaryotes inférieurs sont *Saccharomyces cerevisiae* et *Neurospora crassa*. Les eucaryotes supérieurs sont *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* et *Homo sapiens*.

1.2.2. Seconds acteurs du cycle : les hôtes

Les maladies infectieuses peuvent être strictement humaines (souche du Virus de l'Immunodéficience Humaine HIV-1 M, par exemple) ou strictement animales (fièvre aphteuse), mais nombres d'entre elles sont des zoonoses, c'est à dire transmissibles de l'animal à l'homme et inversement. Les agents infectieux pouvant infecter plusieurs espèces d'hôtes sont qualifiés de **multi-hôtes**. Ils sont plus fréquents que les agents infectieux n'infectant qu'une seule espèce et environ 60 % des pathogènes humains sont zoonotiques [Cleaveland *et al.*, 2001 ; Woolhouse *et al.* 2001]. La maladie de Lyme ou la maladie du sommeil sont des exemples de maladies à transmission vectorielle causées par un agent multi-hôtes. Ces agents sont considérés comme des généralistes, en comparaison avec les spécialistes d'un seul hôte. L'évolution pourrait favoriser les spécialistes [Whitlock, 1996] puisque ces espèces devraient être capables d'évoluer plus rapidement du fait de leur niche écologique restreinte et que les compromis fonctionnels, entre les traits qui permettent l'adaptation à plusieurs espèces et ceux qui favorisent la survie ou la reproduction, devraient réduire la valeur sélective des espèces infectant plusieurs hôtes. Cependant, les agents pathogènes multi-hôtes sont

majoritaires. En effet, il est probable qu'ils soient moins sujets aux extinctions, car leur survie ne dépend pas uniquement d'une seule espèce hôte [Timms and Read 1999] et le fait qu'un agent pathogène puisse être véhiculé par plusieurs hôtes, facilite sa transmission et sa dissémination [Cleaveland *et al.*, 2001]. Dans ce contexte, comprendre comment la capacité à infecter plusieurs hôtes influence l'évolution de la virulence de ces pathogènes et l'épidémiologie des maladies qu'ils provoquent, restent des questions ouvertes [Begon, 2008].

De nombreuses études indiquent que lorsque la **diversité d'hôtes**, qui tient compte de la richesse d'espèces ou nombre d'espèces et de l'abondance de ces espèces, est augmentée ou diminuée dans un milieu donné, le risque d'une maladie infectieuse peut lui aussi augmenter ou diminuer en conséquence [Keesing *et al.*, 2010 ; Ostfeld and Keesing, 2000 ; Ostfeld, 2009 ; Schmidt and Ostfeld, 2001 ; Johnson *et al.*, 2008 ; Johnson and Thielges, 2010]. Si une augmentation de la diversité d'hôtes conduit à augmenter ou diminuer la probabilité que le vecteur rencontre un hôte compétent, il s'agira respectivement d'un effet d'amplification ou de dilution. Par conséquent, le risque maximal lié à la maladie est atteint lorsque la communauté d'hôtes est entièrement composée d'espèces dont la compétence est élevée. Ainsi, un des défis majeurs posé aussi bien par l'émergence de nouvelles maladies que par l'introduction ou l'extinction d'espèces, est de quantifier les **contributions des espèces d'hôtes** à la dynamique de transmission de la maladie. Cette quantification permet d'identifier les mesures clés de prévention à mettre en place afin de réduire la transmission d'agents pathogènes vers certaines espèces d'hôtes comme l'homme, le bétail ou les espèces menacées de disparition [Haydon *et al.*, 2002]. Cependant, elle n'est pas évidente puisqu'il faut prendre en compte, non seulement les caractéristiques propres à l'espèce hôte mais aussi les caractéristiques des populations mises en jeu (Figure 4). Dans ce contexte, on cherche à décrire le réservoir de la maladie qui est constitué d'un groupe d'espèces d'hôtes qui contribue à la persistance de la maladie chez une espèce cible, c'est à dire une espèce d'intérêt dans un contexte particulier [Haydon *et al.*, 2002 ; Viana *et al.*, 2014]. Le réservoir peut être constitué d'espèce(s) de maintenance (taille de population > taille critique de la communauté), qui permettent par leur présence le maintien de l'infection, et d'espèce(s) de non-maintenance (taille des populations < taille critique de la communauté), qui contribuent au réservoir mais dont la présence n'est pas indispensable au maintien de la maladie. L'identification et la caractérisation de ces populations ouvrent des perspectives d'ingénierie des écosystèmes pour mieux contrôler la diffusion des maladies.

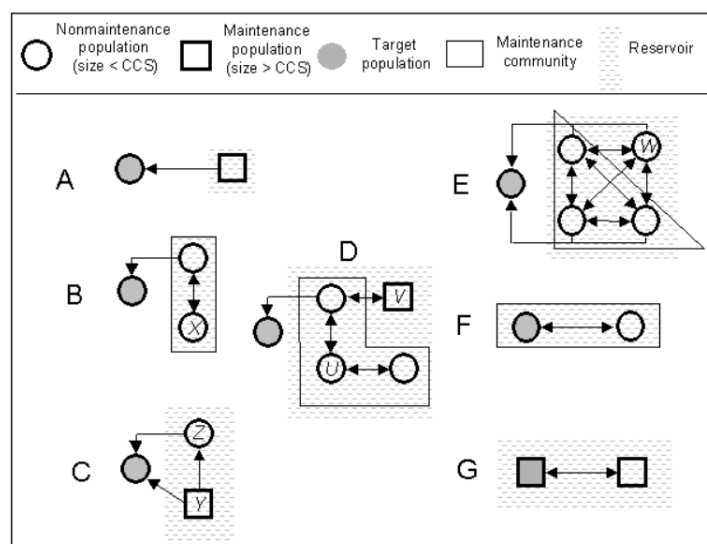


Figure 4 : Les systèmes cible-réservoir [Haydon *et al.*, 2002]

CCS : Critical Community Size. En épidémiologie théorique, la Taille Critique de la Communauté est la taille minimale d'une population fermée dans laquelle un agent pathogène peut persister indéfiniment. Elle permet ici de distinguer les populations de maintenance des populations de non-maintenance. A) La population de maintenance permettant la persistance de l'agent pathogène, transmet un agent pathogène à une population cible. B) Le réservoir est composé de deux populations de non-maintenance connectées, dont une est la source d'infection. Aucune des deux populations ne pourrait constituer un réservoir à elle seule. C) Y est une population de maintenance, mais la transmission peut se produire directement entre Y et la population cible ou par le biais d'une autre population source notée Z. Même si elle n'est pas essentielle à la maintenance du pathogène, Z fait partie du réservoir car elle contribue à la transmission à la cible. D) 4 populations non cibles doivent être incluses dans le réservoir si sa dynamique globale veut être étudiée. L'élimination de l'infection en U ne va pas conduire à l'élimination de l'infection pour la cible parce que V est une population de maintenance indépendante. E) Toutes les populations sont sources. Si W n'est pas nécessaire pour maintenir l'infection, W ne fait pas partie de la communauté de maintenance mais fait partie du réservoir étant une source. F) La cible peut faire partie du réservoir. G) La population cible peut être une population de maintenance.

L'étude des notions de taille critique et de réservoir peut être abordée à l'aide de modèles étudiant la variation du nombre de cas en fonction du temps et éventuellement de l'espace. Les **modèles à compartiments** sont une manière classique d'aborder ce type de problèmes [Anderson and May, 1991]. Ils consistent à diviser la population hôte en autant de compartiments que de statuts immunitaires ou états cliniques et à connecter ces compartiments entre eux par des flux d'individus. Le plus simple des modèles, le modèle SI, n'est composé que de deux compartiments. Dans ce modèle, tous les individus hôtes sensibles (S) auront la même probabilité de passer dans le compartiment des hôtes infectés (I). Si l'on veut prendre en compte le phénomène d'immunité acquise, il faut alors rajouter un

compartiment R (résistants). Des compartiments peuvent être ajoutés au modèle en fonction du degré de précision souhaité et du système étudié comme des maladies à transmission vectorielle par exemple. Une fois la structure du modèle spécifiée, des équations différentielles permettent de le décrire. Ainsi, le modèle SIR peut être décrit par les trois équations suivantes (**Figure 5**):

$$(1) \quad dS/dt = \mu N - (\lambda + \mu)S$$

$$(2) \quad dI/dt = \lambda S - (\gamma + \mu)I$$

$$(3) \quad dR/dt = \gamma I - \mu R$$

où t est le temps continu et $N = S + I + R$ est le nombre total d'individus dans la population, μ le taux de natalité/mortalité, λ la force d'infection (**Intro 1.2.1**) et γ le taux de guérison. La dynamique du système peut être connue en résolvant le système d'équations différentielles *via* des algorithmes de résolution numérique. Tant qu'une épidémie n'est pas sur le déclin, le nombre d'individus dans le compartiment I augmente, c'est-à-dire qu'il y a plus de nouveaux infectés que de nouveaux guéris : $\lambda > \gamma + \mu$. C'est de cela que découle le nombre de reproduction de base R_0 qui correspond au nombre d'infections secondaires suite à l'introduction d'un individu hôte infecté dans une population d'hôtes entièrement constituée d'individus susceptibles [Anderson and May, 1991] et qui équivaut à $R_0 = \beta/(\mu + \gamma)$ avec $\beta = \lambda/I$ (transmission densité dépendante). A partir de tels modèles, différents scénarii de circulation des pathogènes vont ainsi pouvoir être testés et être utilisés pour mieux comprendre les dynamiques observées [Toni *et al.*, 2009] et pour la prise de décisions dans le cadre du contrôle ou la prévention des maladies. C'est pourquoi ces modèles sont un outil de choix pour les épidémiologistes.

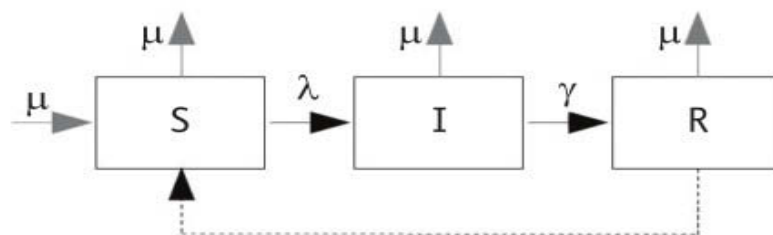


Figure 5 : Schématisation d'un modèle SIR

Les compartiments nommés S, I et R correspondent respectivement aux individus sensibles, infectés et résistants. μ est le taux de natalité/mortalité, λ la force d'infection et γ le taux de guérison.

1.2.3. Troisièmes acteurs du cycle : les vecteurs

Pour les maladies à transmission vectorielle, les agents infectieux utilisent des

Introduction

vecteurs, qui peuvent être indispensables à leur dispersion, pour être véhiculés et transmis d'un individu hôte à un autre. Les vecteurs sont souvent des invertébrés hématophages et parmi eux, on retrouve par exemple des acariens tels que les tiques ou des insectes comme les moustiques. Une infection est transmise par le vecteur après qu'il ait lui-même été infecté, le plus souvent, au cours d'un **repas sanguin** sur un hôte porteur de l'agent pathogène. A la suite de ce repas infectant, l'agent peut se répliquer et/ou se transformer dans le vecteur avant d'être de nouveau transmis lors d'un autre repas de sang. Le plus souvent, la transmission du vecteur à l'hôte s'effectue lors de piqûres. C'est le cas pour les maladies comme le paludisme, le chikungunya ou la maladie du sommeil [Frischknecht *et al.*, 2004 ; Diallo *et al.*, 1999] dont les agents responsables sont transmis lors de piqûres par des moustiques du genre *Aedes* et par les mouches Tsé-Tsé respectivement. Cependant, d'autres modes de transmission du vecteur aux hôtes sont possibles. Par exemple, les trypanosomes responsables de la maladie de Chagas se transmettent *via* les déjections d'une punaise et pénètrent dans l'organisme des individus lorsqu'ils se grattent.

En épidémiologie, les vecteurs sont notamment caractérisés par leur **répartition** et leur **capacité à transmettre** un agent infectieux donné. La capacité prend en compte les conditions du milieu. Elle dépend i) de la compétence du vecteur, c'est à dire de son aptitude à s'infecter sur un hôte vertébré, à assurer le développement de l'agent infectieux et à transmettre cet agent à un autre hôte, ii) du taux de contact vecteur-hôte, lui-même dépendant de la préférence trophique (choix de l'espèce de vertébré pour le repas sanguin) et iii) de l'abondance ou de la densité de vecteurs, ainsi que iv) de la longévité du vecteur ou du nombre de repas qu'il prend durant sa vie. Les variations de compétence et de capacité des vecteurs ont une influence sur la diffusion des agents pathogènes. En effet, lorsqu'il est transmis, par le vecteur à l'hôte, le pathogène peut être soumis à des goulots d'étranglement et subir une diminution de diversité génétique. Par exemple, il a été montré que des arbovirus transmis par des moustiques subissaient des goulots d'étranglement dans différents tissus du moustique conduisant à la transmission d'un nombre de clones à l'animal significativement plus faible que ce qui était initialement présent chez le moustique [Forrester *et al.*, 2012]. Ainsi, l'impact potentiel de la transmission vectorielle sur la diversité et la valeur sélective des pathogènes souligne l'importance d'étudier cet aspect des cycles épidémiologiques.

La **biologie des vecteurs** dicte la dynamique de transmission d'agents pathogènes

qu'ils transmettent. Les vecteurs tels que les moustiques ont généralement des cycles de vie courts pendant lesquels ils prennent plusieurs repas de sang. Ceci leur permet d'avoir des contacts intermittents mais fréquents avec les hôtes sur lesquels ils se nourrissent. Par exemple, la capacité des vecteurs et des hôtes à se déplacer librement, assure une dispersion rapide du virus du Nil Occidental dans des régions contiguës conduisant à des vagues d'infection [Rappole *et al.*, 2006]. En effet, ces épidémies passent par une phase rapide d'introduction dite « silencieuse », puis par une phase de transmission exponentielle et enfin par une phase de décroissance rapide de la maladie, due en partie à la vaccination et/ou au dépeuplement rapide des populations d'hôtes aviaires [Hayes *et al.*, 2005]. Contrairement à cela, les tiques dures ne se nourrissent qu'une seule fois au cours de chaque stade de leur cycle de vie qui peut durer plusieurs années. Elles conduisent généralement à des épidémies persistantes qui se développent lentement [Falco *et al.*, 1995]. Pour les spirochètes responsables de la maladie de Lyme, le vecteur comme les hôtes peuvent servir à maintenir à long terme le pathogène [Tälleklint and Jaenson, 1995], mais la dispersion efficace de l'agent n'est vraiment due qu'aux hôtes. Par exemple, c'est la fixation des tiques *Ixodes* à des hôtes tels que les passereaux qui permet un transport rapide et sur de relatives longues distances de la bactérie [Comstedt *et al.*, 2006]. Cependant, les conditions d'accueil doivent être bonnes à la fois pour les tiques et pour les spirochètes pour qu'un nouveau foyer épidémique puisse s'installer. De plus, les quantités de contact entre les vecteurs et les hôtes conditionnent la prévalence des génotypes dans les populations d'hôtes et donc les opportunités de co-infection, qui peuvent être, comme décrit plus haut, très importantes.

Pour se nourrir avec succès sur un hôte, le vecteur hématophage doit être capable de contourner les défenses de l'hôte (immunitaires ou comportementales, comme le toilettage). Les effets de l'immunité contre les vecteurs affectent la biologie de ceux-ci. Il a été montré par exemple que les défenses des hôtes contre les tiques peuvent conduire à la prolongation de la durée du repas [Schorderet and Brossard, 1993], à une diminution des proportions de tiques qui se gorgent complètement [Girardin and Brossard, 1990] ou à la mort de la tique sur l'hôte. En réponse à ces défenses et pour être capable de mieux exploiter les hôtes, des résultats suggèrent que des phénomènes de **spécialisation des vecteurs aux hôtes** peuvent avoir lieu. Par exemple, dans des populations locales, le choix de l'hôte par la tique *Ixodes ricinus* pourrait être non aléatoire et le résultat de la spécialisation des tiques pour différents types d'hôtes [Kempf *et al.*, 2011]. Cependant, d'autres études sont encore nécessaires pour bien

Introduction

comprendre les processus de spécialisation des tiques pour des espèces d'hôtes et faire le lien avec les agents pathogènes qui sont transmis lors des repas sanguins [McCoy *et al.*, 2013] (Figure 6). Du point de vue de l'homme, dans le cas où les vecteurs seraient spécialistes de certaines espèces d'hôtes, les études basées sur les prévalences d'infection chez la population totale de vecteurs, vont conduire à une sous- ou sur-estimation significative du risque d'exposition à la maladie. C'est pourquoi il s'avère nécessaire d'étudier plus en détail ces processus de spécialisation de vecteurs.

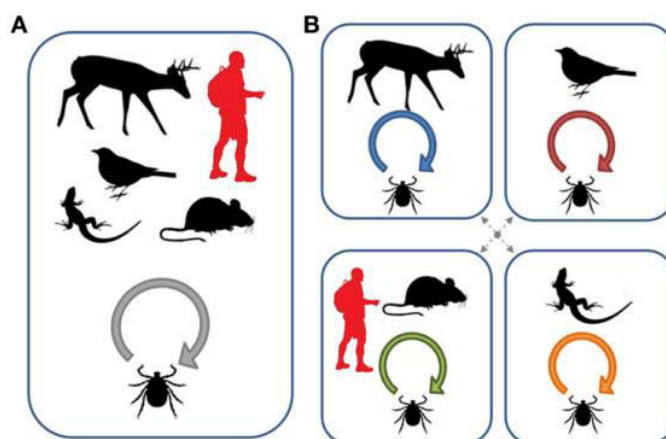


Figure 6 : La circulation des agents pathogènes au sein d'une communauté d'hôte est dépendante de la spécialisation des vecteurs [McCoy *et al.*, 2013]

A) Plusieurs espèces hôtes sont disponibles localement et toutes sont utilisées par le vecteur de la même façon (vecteur généraliste). L'évolution de l'agent infectieux dépendra donc des pressions de sélection imposées par les différents hôtes. B) Chaque espèce hôte vertébré est exploitée par une population de vecteurs différente, ce qui conduit localement à des cycles de transmission de la maladie indépendants ou semi-indépendants (flèches en pointillés). L'évolution de l'agent infectieux dépendra donc du degré de spécificité des différentes populations de vecteurs et de sa capacité à exploiter les différents hôtes après sa transmission.

2. La biologie moléculaire et la génétique des populations au service de l'étude des maladies vectorielles

Comprendre comment circulent et évoluent les maladies à transmission vectorielle, devient un des challenges majeurs de l'épidémiologie d'aujourd'hui, notamment pour fournir les informations nécessaires à la mise en place de stratégies de prévention et de gestion des événements de maladies émergentes à transmission vectorielle. Ces études nécessitent une meilleure prise en compte de la diversité des agents pathogènes et des interactions entre les agents pathogènes, les hôtes (hommes, animaux domestiques ou animaux sauvages), les vecteurs et l'environnement. Ceci implique d'être confronté à la complexité et à la variabilité de ces systèmes. Dans ce contexte, l'application des méthodes classiques de la **biologie**

moléculaire et de **biologie évolutive**, peuvent permettre de répondre à des questions d'ordre épidémiologique [Foxman and Riley, 2001].

2.1. Intégration des concepts phylogénétiques dans les approches d'épidémiologie moléculaire

Historiquement, les méthodes de **biologie moléculaire** ont été utilisées dans l'objectif de caractériser, des isolats pour les comparer, puis valider ou réfuter leur lien épidémiologique. La caractérisation du polymorphisme nucléotidique de ces isolats est passée par le développement et/ou l'utilisation de différentes méthodes de biologie moléculaire dont le pouvoir résolutif et/ou la reproductibilité sont plus ou moins limitées. Tout d'abord, des profils de polymorphisme de longueur de fragments de restriction (RFLP, Restriction Fragment Length Polymorphism) couplée à la PFGE (Pulsed Field Gel Electrophoresis) peuvent être établis pour établir les génotypes auxquels on est confronté. D'autres méthodes qui ont également vu le jour, peuvent au préalable faire appel à une étape d'amplification de différents marqueurs. Par exemple, l'analyse des profils de digestion enzymatique de l'ADN par PCR-RFLP ou AFLP (Amplified Fragment Length Polymorphism) a permis une caractérisation des isolats à une échelle inter ou intra-spécifique.

Cependant, l'avènement des méthodes de séquençage a permis d'ouvrir de nouvelles perspectives d'études en augmentant considérablement la résolution de l'information obtenue. Au-delà de l'identification des génotypes, il s'agit notamment de caractériser de nouveaux isolats d'agents pathogènes i) en établissant leurs relations de parenté avec d'autres isolats déjà connus de la même ou d'autres espèces et ii) en obtenant des informations quant à leur évolution. Concernant la diversité bactérienne, la plupart de nos connaissances sont issues du génotypage de nombreuses souches grâce au séquençage de plusieurs loci (MLST Multi Locus Sequence Typing) [Maiden *et al.*, 1998]. La **MLST**, reproductible et sans ambiguïté, permet un typage définitif des souches qui peut être facilement partagé entre les laboratoires notamment *via* des bases de données. Les marqueurs utilisés dans le cadre d'études de MLST sont des gènes de ménage soumis à des pressions de sélection purificatrices à cause desquelles les mutations délétères sont rapidement éliminées, ne laissant qu'une majorité de polymorphismes quasi-neutres dans les populations. Ces données de séquences apportent une nouvelle dimension dans l'élucidation des relations de parenté des souches à l'échelle inter et intra-spécifique et, plus le nombre de loci séquencés augmente, plus la résolution sera

Introduction

importante [Maiden *et al.*, 2013] (Figure 7).

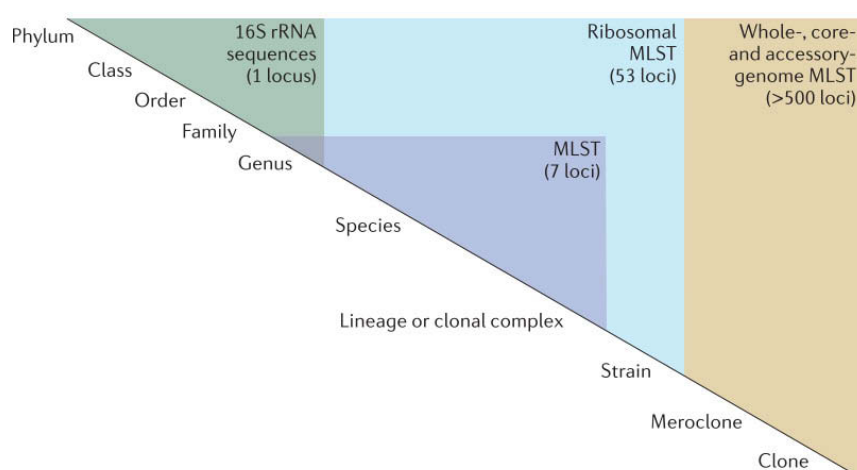


Figure 7 : La résolution est dépendante du type et du nombre de marqueurs utilisés [Maiden *et al.*, 2013]

A l'ère du séquençage haut-débit à bas-coûts, il devient également de plus en plus fréquent d'obtenir des **séquences de génomes entiers ou partiels**, bien que l'isolement et la mise en culture du pathogène étudié soient nécessaires mais pas toujours possibles. Pour ces études, le nombre de souches que l'on caractérise est souvent bien plus restreint que ce que permet la MLST, mais les études génomiques permettent d'obtenir des niveaux de résolution encore plus fins pour les analyses de diversité génétique et de phylogénie.

Les **arbres phylogénétiques** construits à partir de séquences homologues permettent une représentation schématique des distances évolutives qui séparent les échantillons qui composent l'arbre à un instant donné. Lorsque l'on possède un échantillonnage à différentes dates, souvent plusieurs années, les phylogénies peuvent être utilisées pour calibrer des horloges moléculaires pour les loci étudiés, pour estimer des taux d'évolution pour les séquences ou dater les ancêtres communs de groupes d'individus [Drummond *et al.*, 2002 ; Didelot and Falush, 2007] *via* l'optimisation des paramètres de fonctions de vraisemblance, par exemple. Des travaux sur le virus de la maladie hémorragique épizootique (EHDV) vectorisé par des moustiques, ont permis d'estimer un taux d'évolution sur 40 ans de $4,75 \times 10^{-4}$ substitution/site/an du sérotype EHDV-2 et de dater l'ancêtre commun des souches de ce sérotype aux Etats-Unis à une centaine d'années [Biek, 2007]. Cependant, ces méthodologies, sont complexes à aborder d'un point de vue statistique et il peut être difficile d'approximer au mieux la distribution *a posteriori* des paramètres d'intérêts à partir des données et d'une distribution *a priori*.

Pour appréhender plus facilement ces modèles complexes, il est possible d'avoir une

approche par simulation de jeux de données en fonction de valeurs de paramètres (recombinaison, migration, mutation...) variables, et à les comparer aux données observées *via* une **Approche Bayésienne Approchée** (ou Approximate Bayesian Computation, ABC) [Beaumont, 2010 ; Bertorelle *et al.*, 2010] (Figure 8). Les priors sont déterminés de sorte qu'ils permettent d'explorer un espace de valeurs de paramètres vaste dans lequel vont se situer les valeurs de paramètres que l'on va inférer. L'impact des différents paramètres sur les données simulées, est mesuré grâce au calcul de statistiques descriptives de génétique des populations. Ces mesures permettent de calculer une distance euclidienne multivariée entre les données observées et les données simulées et ainsi de mesurer l'adéquation simulées-observées. La distance prend en compte les différentes statistiques résumées qui ont été choisies. Lorsque la distance est plus faible qu'une distance de tolérance, définie de telle sorte qu'une petite fraction des simulations soit acceptée, les paramètres sont acceptés. C'est à partir des distributions des paramètres acceptés que les inférences sur les données observées sont réalisées. Par rapport aux méthodes classiques d'estimations par maximisation de fonctions de vraisemblance, ces approches bayésiennes sont moins complexes d'un point de vue statistique [Beaumont and Rannala, 2004] et l'ABC offre davantage d'opportunités d'aborder d'une manière assez simple des questions pour lesquelles des outils n'ont pas encore été implémentés.

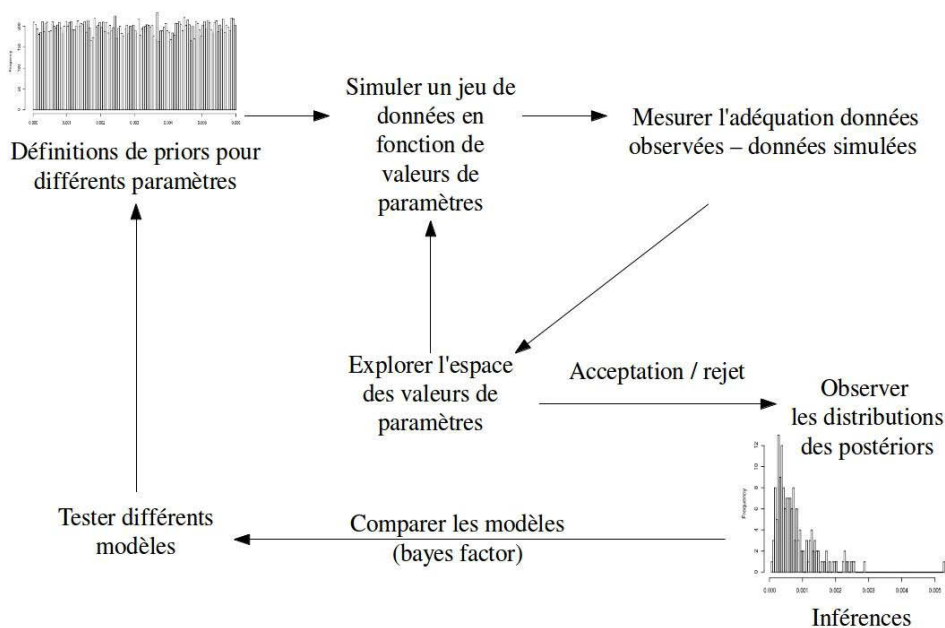


Figure 8 : Principe de l'approche Bayésienne et inférences

2.2. Apport de la génétique des populations au développement des approches phylodynamiques

Pour permettre de comprendre les **dynamiques de population** à l'aide de données génétiques et *via* la prise en compte de plusieurs processus évolutifs et épidémiologiques en même temps, des modèles qui utilisent la **théorie de la coalescence** peuvent être construits et étudiés à l'aide d'une approche ABC. Ce type de modèles, classiquement utilisés en génétique des populations, peut en effet apporter des outils pour résoudre des problèmes concrets rencontrés en épidémiologie en s'intéressant aux questions qui concernent la diversité des pathogènes dans une population. A travers la généalogie d'échantillons, on va chercher à identifier quels sont les forces sélectives ou les processus épidémiologiques qui font évoluer leurs séquences. En effet, en plus des informations qu'ils apportent sur les liens de parenté, les polymorphismes observés sont la conséquence de quatre forces évolutives majeures qui contribuent à la modification des fréquences alléliques des séquences nucléotidiques. La première est la mutation qui génère de nouveaux allèles et la seconde est la sélection qui conduit à l'augmentation en fréquence d'allèles qui confèrent un avantage aux individus. La troisième force est la dérive génétique (stochasticité due à l'échantillonnage aléatoire des parents dans la population) qui est d'autant plus forte que la taille de la population N est petite et qui peut conduire à la fixation d'allèles mutants aléatoirement. La dernière force est la migration de gènes ou flux génique. Le flux génique est le mouvement des gènes entre différents pools génétiques [Slatkin, 1985], ayant pour résultat final l'homogénéisation des allèles entre populations. Plus spécifiquement, de nouveaux variants bactériens peuvent être générés grâce à la recombinaison entre segments d'ADN homologues ou l'intégration d'ADN exogène par transformation et ou par conjugaison [Maynard Smith *et al.*, 1993 ; Matic *et al.*, 1995]. Ainsi, l'adéquation de scénarii faisant intervenir des **variations démographiques de populations**, des structurations de population comme la **spécialisation** liée à un effet d'hôte, ou encore la **recombinaison homologue** [Hudson, 1991 ; Wiuf and Hein, 2000 ; Nordborg, 2000] aux données observées pourra être testée (Figure 9). Cependant, d'autres outils, dont pour certain une brève description est donnée ci-dessous, sont disponibles pour étudier ces différents processus de façon indépendante.

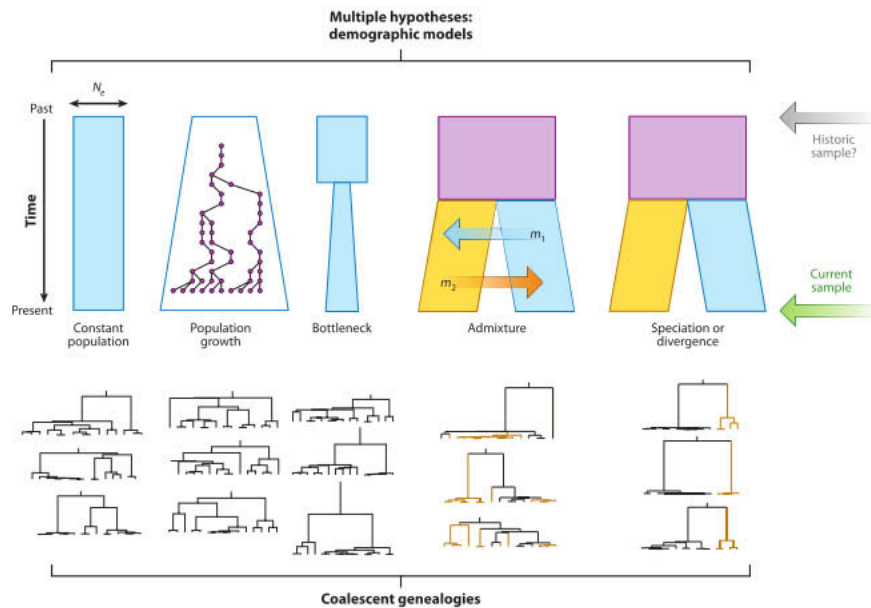


Figure 9 : Illustration de différents scénarii évolutifs pouvant être testés *via* une analyse de coalescent [Grünwald and Goss, 2011]

Une population d'individus (échantillons actuels ou historique) est obtenue. Pour chaque individu, plusieurs loci génétiques sont caractérisés et sujets à une analyse de coalescence pour obtenir des généalogies dans le but d'inférer les processus évolutifs qui ont conduit aux profils des données observées. Différents modèles démographiques peuvent être testés (taille de la population constante, expansion ou goulot d'étranglement) tout comme les événements de migration, de recombinaison ou les phénomènes de divergence.

Dans le cadre de l'étude des dynamiques de population de pathogènes, nous avons vu qu'il pouvait s'avérer intéressant de prendre en compte les **variations démographiques** des populations (**Intro. 1.2.3**). Ces événements entraînent des changements dans l'intensité de la dérive génétique s'exerçant sur la population et, de ce fait l'équilibre entre la mutation et la dérive n'est plus conservé. Ceci affecte la distribution des génotypes observés dans une population par rapport à ce qui aurait été observé dans le cadre d'une évolution à l'équilibre. Un goulot d'étranglement est un bon exemple d'une transition démographique susceptible de modifier les généalogies d'un échantillon de loci. Rechercher des écarts au modèle neutre standard est une voie pour réaliser des inférences sur l'histoire évolutive, sélective ou démographique d'une population. Tajima [1989] a proposé un des premiers tests d'écart à l'équilibre basé sur la diversité moléculaire d'un échantillon. Le principe de ce test est de comparer les estimations du paramètre de mutation $\theta = 2N\mu$ (pour les haploïdes), avec μ le taux de mutation par site nucléotidique et par génération et N la taille de la population, obtenues i) à partir du nombre de sites polymorphes S (θ_S) et ii) à partir du nombre moyen de

Introduction

différences entre deux séquences (θ_π). Pour les loci qui évoluent sous sélection neutre, en cas d'expansion démographique de nombreux allèles rares ou à faible fréquence sont maintenus, ce qui entraîne peu de modification de θ_π mais une augmentation de θ_S , si bien que l'on s'attend à avoir des valeurs négatives de D . Dans le cas d'une diminution de la taille de population, les allèles de fréquences intermédiaires sont favorisés et le nombre de sites polymorphes diminue d'autant plus vite que la taille de l'échantillon est petite. Ces phénomènes ont beaucoup d'effet sur θ_π mais relativement peu sur θ_S . D sera donc positif. Dans le cas d'un goulot d'étranglement on observe successivement les phases de contraction et d'expansion de la population. Au cours des générations D sera donc d'abord positif puis négatif (Figure 10).

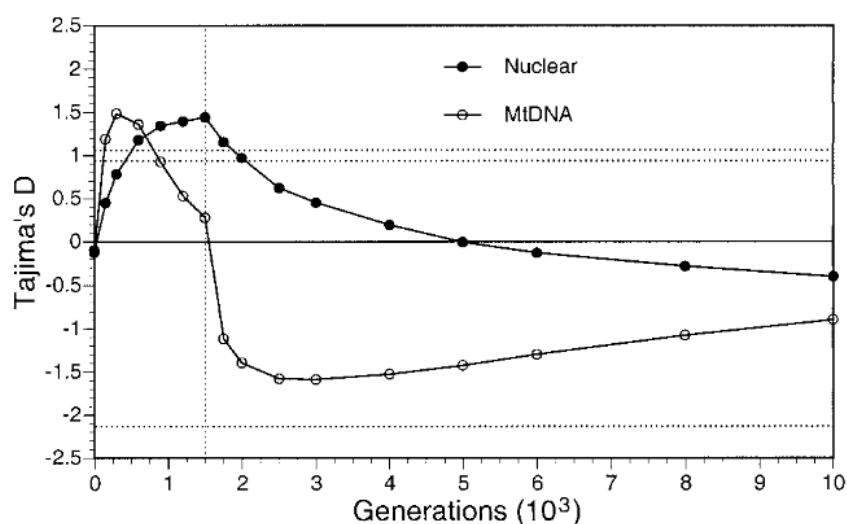


Figure 10 : Evolution de la valeur du D de Tajima au cours des générations après un goulot d'étranglement [Fay and Wu 1999]

Le long des génomes, certaines régions peuvent être soumises à des **pressions de sélection** particulières. Dans le cadre de l'étude de maladie infectieuse, il est alors intéressant de pouvoir détecter des gènes pouvant être liés à l'infection, par exemple, qui sont soumis à des pressions de sélection diversifiantes ou balancées. L'utilisation du test de Tajima [1989] ou de la comparaison du nombre de mutations synonymes par rapport aux mutations non synonymes, sont des outils pour détecter ces régions spécifiques le long des génomes. De plus, comme décrit précédemment, les pressions de sélection exercées par les hôtes peuvent conduire à l'émergence et la maintenance de variants adaptés à un seul hôte ou à un spectre d'hôte restreint. Ces phénomènes de **spécificité d'hôtes** sont une préoccupation majeure dans

l'étude des dynamiques des maladies vectorielles multi-hôtes. Ils peuvent générer de la différenciation génétique et conduire à un isolement reproductif en absence de brassage génétique) Ainsi, l'utilisation de mesures de différenciation tel que l'indice de fixation F_{ST} sont également un bon moyen pour identifier des gènes candidats qui pourraient être sous sélection [Beaumont, 2005]. En effet, des degrés de différenciation faibles (~ 0) ou élevés (~ 1) sont la conséquence i) des pressions de sélection notamment immunes qui s'exercent sur certaines souches, mais également ii) de l'isolement géographique ou écologique entre les souches qui composent les populations étudiées, iii) des flux d'individus entre les populations par migration, et iv) de la fréquence des événements de recombinaison génétiques (Figure 11).

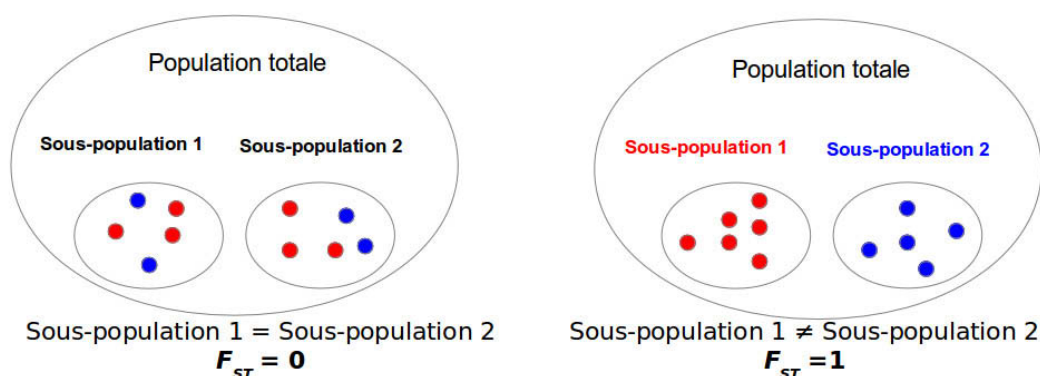


Figure 11 : Différenciation génétique, exemple de comportement de l'indice de fixation F_{ST}

Lorsque les fréquences alléliques sont proches dans deux sous-populations (Sous-population 1 et Sous-population 2), alors les deux sous-populations ne sont pas différenciées et l'indice de fixation calculé sera proche de 0 (à gauche). Si les fréquences des allèles dans deux sous-populations, sont totalement différentes alors les deux sous-populations sont différenciées et l'indice de fixation calculé sera proche de 1 (à droite).

L'évaluation de l'impact de la **recombinaison homologue** sur la diversité de souches bactériennes est, une question intrigante de la biologie évolutive posée par la disponibilité à grande échelle de données de séquence. Tout d'abord, des outils phylogénétiques adaptés à leur étude ont été développés [Huson and Bryant, 2006]. En effet, les réseaux permettent une visualisation des ambiguïtés phylogénétiques, grâce à des réticulations entre les branches, qui peuvent être dues à des événements de transfert de gène horizontaux ou de recombinaison. D'autre part, on peut effectuer des mesures de **déséquilibre de liaison** qui sont dépendantes du taux de recombinaison [Spratt *et al.*, 2001] (Figure 12). Lorsqu'il n'y a pas de recombinaison, l'association des allèles à deux loci différents ne se fait pas de manière aléatoire au sein de la population. Ils sont dits en déséquilibre de liaison [Lewontin, 1964]. Il est estimé que pour éliminer totalement le déséquilibre de liaison au sein d'une population bactérienne, le taux de recombinaison par locus doit être 20 fois supérieur au taux de mutation

Introduction

[Hudson, 1994]. Cependant, Maynard Smith et collaborateurs ont mis en évidence que l'impact des événements de recombinaison peuvent être masqués par des différences dans le succès de la reproduction de génotypes recombinants lorsque l'on réalise un sous-échantillonnage de la population totale [Maynard Smith *et al.*, 2000] (Figure 13). Dans cette structure de population dite épidémique, les différences de fréquences des génotypes recombinants conduisent à une augmentation du déséquilibre de liaison malgré des taux de recombinaison qui peuvent être élevés. Ainsi, l'estimation des taux de recombinaison homologue impactant une population reste un problème complexe et, dans l'interprétation de celui-ci, il est important de prendre en compte l'impact de l'échantillonnage réalisé.

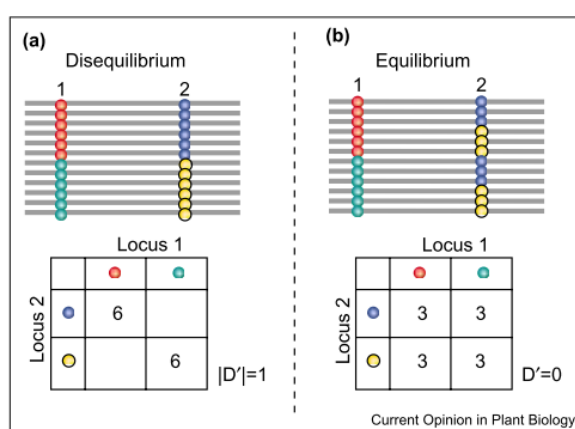


Figure 12 : Déséquilibre de liaison et recombinaison [Rafalski, 2002]

a) Tous les individus qui ont l'allèle rouge au locus 1 ont l'allèle bleu au locus 2. On est en situation de déséquilibre de liaison ($D'=1$) ce qui correspond à une situation où il n'y a pas eu de recombinaison. b) Les individus qui ont l'allèle rouge au locus 1 ont l'allèle bleu et l'allèle jaune en proportion égale au locus 2. On est en situation d'équilibre de liaison qui a pu être provoqué par un événement de recombinaison au locus 2.

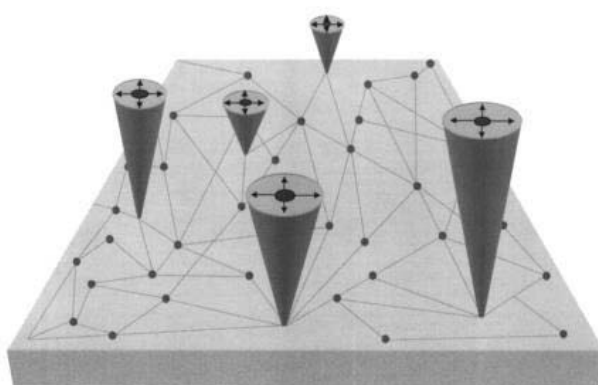


Figure 13 : Structure épidémique [Maynard Smith *et al.*, 2000]

La population est composée de deux parties. La population d'« arrière-plan » est composée d'un nombre de génotypes à faible fréquence qui recombinent fréquemment. Parmi ces génotypes, quelques-uns dont le succès de la reproduction est important, vont se retrouver à haute fréquence (clones). La diversification de ces clones est due à la recombinaison mais également à la mutation (représentée par les flèches).

Le développement récent d'approches permettant une intégration de plus en plus fine des processus décrits au-dessus, est à l'origine de l'intérêt grandissant pour le domaine de la **phylodynamique** qui cherche à comprendre comment les processus d'évolution et épidémiologiques interagissent pour façonner les phylogénies des agents pathogènes étudiés. Depuis quelques années ces approches ont permis de grandes avancées dans l'étude des épidémies de virus à évolution rapide et. plus récemment, ont fourni des résultats prometteurs pour des agents pathogènes bactériens à évolution plus lente [Voltz *et al.*, 2013].

Par rapport aux techniques de laboratoire et aux méthodes d'analyses existantes, on peut envisager d'utiliser de nouvelles sources d'information. En effet, jusqu'à présent on considèrerait qu'il n'y avait que des mono-infections, cependant ce n'est très probablement pas toujours le cas et les **co-infections** peuvent être plus ou moins fréquentes. En lien avec la dynamique de circulation des maladies, elles peuvent avoir une importance capitale dans la dynamique (compétition, facilitation)(**Intro. 1.2.1**) ou l'évolution (par échange de matériel génétique) des pathogènes étudiés. Les méthodes classiques de PCR-RFLP permettent de distinguer les individus co-infectés de ceux qui ne le sont pas. Les méthodes de MLST qui font appel à un séquençage Sanger, peuvent également permettre de connaître les loci pour lesquels plusieurs allèles étaient présents dans l'échantillon pour lequel on réalise un séquençage puisque de multiples pics pourront apparaître sur les chromatogrammes de séquences. Cependant ces méthodes ne permettent pas d'identifier proprement les génotypes et donc d'obtenir une information précise qualitative et quantitative des niveaux de co-infections des individus. Le développement de nouvelles approches pourrait permettre de tirer un meilleur parti des données informatives contenues dans la présence de co-infections.

3. Modèle biologique : la borréliose de Lyme

La compréhension du fonctionnement et de l'évolution des maladies à transmission vectorielle peut donc passer par la mise en place d'études d'épidémiologie moléculaire et l'utilisation d'outils de génétique des populations. Dans le cadre de ma thèse, nous avons choisi l'exemple d'une maladie vectorielle, la maladie de Lyme. Cette maladie est causée par les bactéries du complexe d'espèces *B. burgdorferi* s.l. et est transmise par les tiques à de nombreux hôtes vertébrés. En France, l'Institut de Veille Sanitaire inclut la borréliose de Lyme parmi les 20 maladies à surveiller en priorité car elle est la zoonose non alimentaire

Introduction

présentant la plus forte incidence chez l'homme. Cette maladie est considérée comme émergente et potentiellement grave [Steere *et al.*, 2004]. Aux Etats-Unis, de nouvelles estimations du Centre pour le Contrôle et la Prévention des Maladies (CDC), porterait le nombre de nouveaux cas de maladie de Lyme à minima à 300 000 et jusqu'à un million par an [Stricker and Johnson, 2014]. Ainsi, la description de la diversité génétique des pathogènes du genre *Borrelia*, la caractérisation des moteurs d'évolution de leur génomes, tout comme l'identification et la compréhension des cycles épidémiologiques de la bactérie sont d'un intérêt majeur.

3.1. Le pathogène

En 1972, plusieurs dizaines d'enfants et une dizaine d'adultes ont été atteints d'arthrite rhumatoïde dans l'Est du Connecticut. Une surveillance est alors organisée dans des communautés proches de la petite ville de Lyme [Steere *et al.*, 1977]. Après examen clinique de nombreux habitants et évaluation du contexte épidémiologique, on conclut à une maladie vectorielle transmise par un arthropode et nommée arthrite de Lyme. La présence de lésions particulières précédant l'apparition des premiers symptômes d'arthrite chez un quart des patients est similaire aux manifestations cliniques observées plus tôt en Europe : l'érythème chronique migrant, supposé être transmis par **la tique** *I. ricinus* [Lipschütz, 1913]. Il faut attendre les années 80 pour que des spirochètes soient détectés dans la tique *Ixodes scapularis* [Burgdorfer *et al.*, 1982] (Figure 14). Burgdorfer et collaborateurs [Burgdorfer *et al.*, 1982] remarquent que les spirochètes induisent une réaction immunitaire chez les patients atteints par cette arthrite de Lyme, désormais nommée maladie de Lyme. La confirmation viendra juste après, par l'isolement de spirochètes depuis le sang de patients atteints [Benach *et al.*, 1983] et la découverte des spirochètes chez la tique *I. ricinus* en Europe [Burgdorfer *et al.*, 1983]. En 1984, finalement, ce spirochète transmis par les tiques du genre *Ixodes* est taxonomiquement identifié et nommé ***Borrelia burgdorferi*** [Hyde and Johnson, 1984].

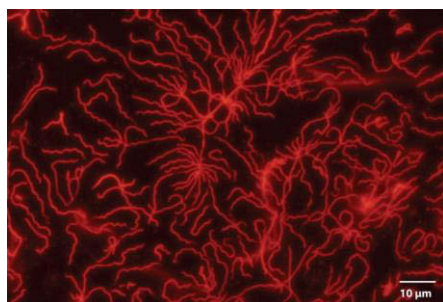


Figure 14 : Spirochètes du genre *Borrelia* [Brisson *et al.*, 2012]

Depuis ces découvertes, les efforts d'échantillonnage liés à la maladie ont été de plus

en plus conséquents et de **nombreuses espèces** ont été décrites. La première est *B. garinii* [Postic *et al.*, 1990 ; Baranton *et al.*, 1992], puis viennent *B. afzelii* [Baranton *et al.*, 1992 ; Canica *et al.*, 1993] et *B. valaisiana* [Wang *et al.*, 1997]. La notion de *B. burgdorferi* sensu stricto (s.s.) pour l'espèce et de *B. burgdorferi* sensu lato (s.l.) pour l'ensemble du complexe de génoespèces, est alors adoptée. Ainsi, le complexe *B. burgdorferi* s.l. contient aujourd'hui 19 génoespèces connues à travers le monde. Six d'entre elles sont avérées pathogènes pour l'homme : *B. burgdorferi* s.s., *B. afzelii*, *B. garinii* et *B. bavariensis* mais également *B. spielmanii* et *B. lusitaniae* dans une moindre mesure. Deux autres génoespèces, *B. bissettii* et *B. valaisiana* sont soupçonnées d'être aussi pathogènes, mais de nouvelles recherches pour le démontrer sont nécessaires [Baranton and De Martino, 2009]. Les espèces *B. americana*, *B. andersonii*, *B. californiensis*, *B. kurtenbachii* aux États-Unis ainsi que *B. carolinensis* en Europe comme aux États-Unis et finalement *B. japonica*, *B. tanukii*, *B. turdi*, *B. sinica* et *B. yangtze* en Asie n'ont jamais été isolées depuis un patient. Finalement, une dernière génoespèce a été proposée en Finlande : *B. finlandensis* [Casjens *et al.*, 2011a].

La borréliose de Lyme est surtout présente dans l'**hémisphère nord** [Kunterbach *et al.*, 2006] (Figure 15) : elle est recensée en Amérique du Nord, en Europe et en Asie. Sa présence est également démontrée en Afrique du nord; en Tunisie, au Maroc et en Algérie. Aux États-Unis, la maladie est surtout présente dans l'Est du pays (Maine, Floride, Minnesota) mais aussi sur la côte Ouest (Californie et état de Washington). Toute l'Europe est touchée, avec une limite jusqu'à la Norvège au Nord et elle s'étend ensuite vers l'est, jusqu'au Japon. Cependant, la répartition de la maladie est étroitement liée à la répartition des vecteurs tiques responsables de leur transmission.

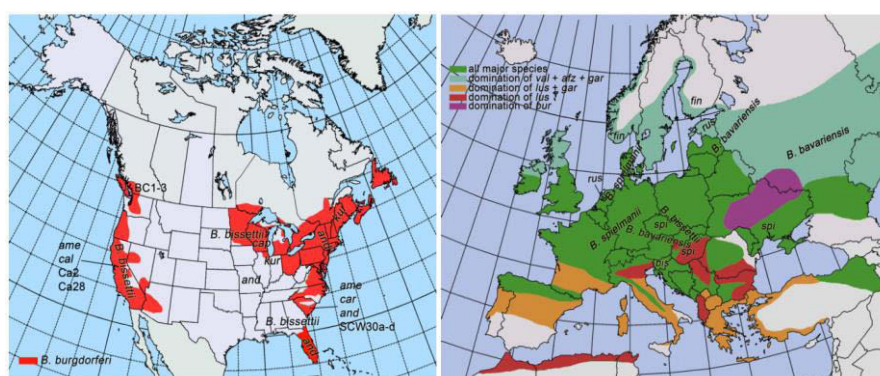


Figure 15 : Distribution géographique de *B. burgdorferi* s.l. aux États-Unis (à gauche) et en Europe (à droite) [Franke *et al.*, 2013]

Les génoespèces majeures (major species) correspondent à *B. afzelii* (afz), *B. burgdorferi* s.s. (bur), *B. garinii* (gar), *B. lusitaniae* (lus) et *B. valaisiana* (val).

Introduction

3.2. Les hôtes

La maintenance de *B. burgdorferi* s.l. dans les zones endémiques requiert la présence d'espèces hôtes réservoirs. Il s'agit notamment des oiseaux comme les passereaux ou les faisans et des petits rongeurs tels que les campagnols, les mulots, les rats, les écureuils ou encore les loirs et lérots. Les grands ongulés, comme le cerf et le chevreuil, ne sont, en revanche, pas considérés comme des réservoirs de la maladie de Lyme [Telford *et al.*, 1988 ; Tälleklint and Jaenson, 1994]. Malgré de nombreuses études, le rôle des différentes espèces hôtes intervenant dans le cycle des génoespèces de *B. burgdorferi* s.l. n'est pas encore très bien défini. Cependant, il semblerait que les espèces européennes de *Borrelia* tendent à se spécialiser pour certains taxons vertébrés [Kurtenbach *et al.*, 2006] puisque des **associations spécifiques entre hôtes réservoirs et espèces de *B. burgdorferi* s.l.** ont été observées. *B. burgdorferi* s.s. et *B. bavariensis* sont associées aux petits rongeurs [Kurtenbach *et al.*, 1998 ; Margos *et al.*, 2009] et aux écureuils roux [Humair and Gern, 1998] tout comme *B. afzelii* souvent associée aux mulots sylvestres et aux campagnols roussâtres [Hu *et al.*, 1997]. *B. garinii* et *B. valaisiana* sont associées aux oiseaux et tout particulièrement aux passereaux [Olsen *et al.*, 1995 ; Humair *et al.*, 1998 ; Kurtenbach *et al.*, 1998 ; Hanincova *et al.*, 2003 ; Dubska *et al.*, 2009]. *B. spielmanii* est associée aux lérots ainsi qu'aux hérissons [Richter *et al.*, 2004 ; Skuballa *et al.*, 2007] tandis que *B. lusitaniae* est associée aux lézards [Clark *et al.*, 2005]. Dans ce contexte, une des questions qui se posent concerne l'existence d'un niveau d'association encore plus fin, c'est à dire à l'échelle des génoespèces de *Borrelia*, aux différentes espèces d'hôtes.

Tous les hôtes réservoirs n'ont pas, en revanche, la même **contribution** à la dynamique de transmission de *B. burgdorferi* s.l.. Une immunité spécifique contre les tiques et/ou les *Borrelia* propre à chaque espèce hôte peut être observée ce qui modifie la compétence de réservoir de ces espèces [Humair *et al.*, 1999]. Par exemple, les campagnols ont une charge en tiques moins importante que les mulots et les tiques qui se nourrissent sur les campagnols ont un succès de métamorphose inférieur à celles qui se sont nourries sur les mulots. Il semble néanmoins que les campagnols génèrent plus de tiques infectées que les mulots et que leur contribution à l'infection des tiques est donc plus importante [Kurtenbach *et al.*, 1994 ; Pérez, 2012]. Les modèles développés pour estimer la contribution des hôtes au risque lié à la maladie de Lyme sont principalement basés sur le calcul du R_0 . Par exemple, en utilisant un jeu de données sur l'infestation en tiques de petits rongeurs, d'écureuils et

d'oiseaux en Grande-Bretagne, ainsi que des paramètres issus de la littérature, les importances relatives des différentes espèces d'hôtes dans la composante du R_0 de la maladie ont été comparées [Randolph and Craine, 1995]. Cependant, ce type de modèle ne permet pas d'étudier de façon simultanée des infections par différentes souches de *Borrelia*. C'est pourquoi, d'autres types de modèles ont également fait leur apparition. Par exemple, celui développé par Brisson et collaborateurs [2008] utilise l'assignation de différents génotypes d'une des espèces de l'agent responsable de la maladie de Lyme à différentes espèces hôtes pour estimer la proportion de tiques qui se sont nourries sur chaque espèce d'hôte du système étudié. La prise en compte des taux de transmission permet d'augmenter le pouvoir prédictif de ces modèles puisqu'ils pourront être utilisés notamment pour comprendre l'influence d'un hôte donné sur la fréquence d'un génotype de *B. burgdorferi* s.l.. Cependant, les résultats fournis par ces modèles vont être dépendants de la qualité des estimations de ces taux de transmissions, calculés à partir de données de la littérature ou d'expériences [Brisson *et al.*, 2006 ; Brisson *et al.*, 2008]. Des efforts sont à fournir pour le développement de nouvelles méthodologies d'estimation des contributions des hôtes, indépendantes de la qualité des estimations d'autres paramètres.

3.3. Le vecteur tique

La répartition des différentes génoespèces du complexe *B. burgdorferi* s.l. qui circulent dans l'environnement est évidemment étroitement liée à la distribution géographique des tiques du genre *Ixodes*. Au sein de ce genre, différentes espèces sont capables d'acquérir et/ou transmettre la bactérie ; cependant *I. scapularis* et *Ixodes pacificus* aux Etats-Unis, *I. ricinus* en Europe et *Ixodes persulcatus* en Eurasie sont les vecteurs les plus compétents (Figure 16). Les tiques *I. ricinus* s.l., possèdent **trois stades de développement** actifs : larve, nymphe et adulte. La tique doit prendre un repas sanguin obligatoire pour passer d'un stade à l'autre, à l'exception de l'adulte mâle. C'est lors de ces différents repas sanguins que les tiques acquièrent la bactérie. Cependant, seules les nymphes et les adultes femelles sont capables de transmettre la bactérie à l'hôte sur lequel elles se nourrissent (Figure 17).

Introduction

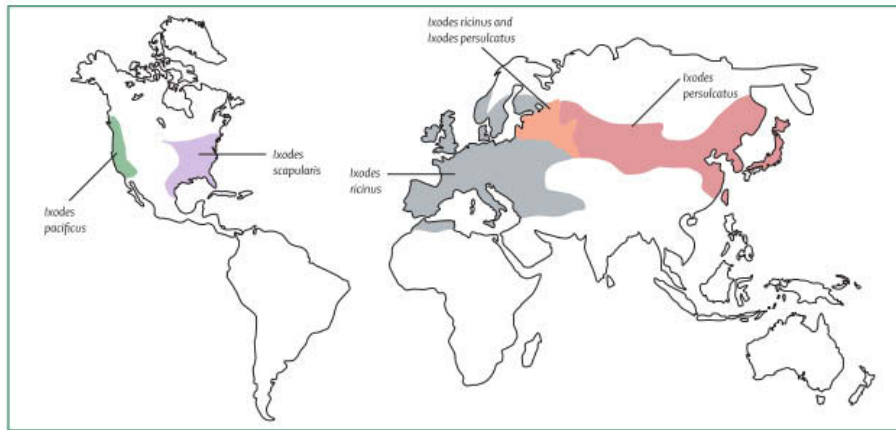


Figure 16 : Distribution géographique de *I. ricinus* s.l. [Stanek *et al.*, 2012]

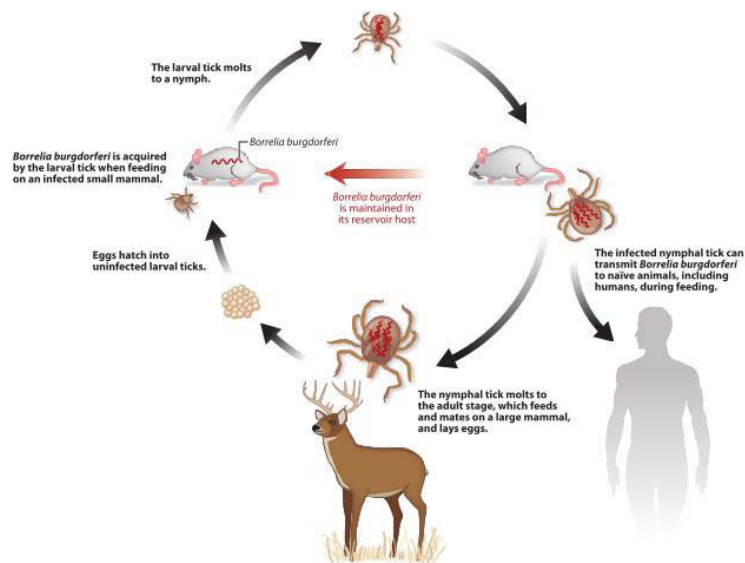


Figure 17 : Cycle de vie des tiques *I. ricinus* associé à celui de *B. burgdorferi* s.l. [Brisson *et al.*, 2012]

Les spirochètes sont acquises par les larves lorsqu'elles se nourrissent sur le premier hôte vertébré (classiquement de petits mammifères ou des oiseaux). Les larves muent ensuite en nymphes qui vont pouvoir transmettre la bactérie en se nourrissant sur un second hôte. Les nymphes muent alors vers le stade adulte. Les femelles vont alors prendre un troisième repas de sang. Tous les stades de développement de la tique peuvent se nourrir sur l'homme (considéré comme hôte accidentel), mais ce sont les nymphes qui sont responsables de la majorité des cas de borréliose de Lyme.

Bien que le stade de développement de la tique qui présente la plus grande prévalence d'infection par la bactérie soit le stade adulte, puisque les infections s'accumulent au cours des différents repas, ce sont les **nymphes** qui constituent le stade le plus important dans le cadre de la transmission de la maladie à l'homme [Piesman and Gern, 2004] (Figure 17). En effet, les nymphes sont beaucoup plus nombreuses que les adultes et passent plus facilement inaperçues du fait de leur petite taille, ce qui leur laisse le temps de transmettre la bactérie.

En dehors d'un repas sanguin sur un hôte infecté, les tiques peuvent s'infecter de deux

autres façons. Tout d'abord, l'infection peut se faire *via* la transmission transovarienne de *B. burgdorferi* s.l., c'est-à-dire la transmission des adultes femelles aux œufs. Cette transmission est cependant considérée comme négligeable car probablement peu fréquente [Bellet-Edimo *et al.*, 2005] (Figure 17). Ensuite la tique peut acquérir les *Borrelia* lors de **co-repas**. La transmission par co-repas a lieu lorsqu'une tique s'infecte au moment du gorgement sur l'hôte par transmission locale de *B. burgdorferi* s.l. à partir d'une autre tique infectée se trouvant à côté d'elle [Gern and Rais, 1996], sans passage de l'infection par l'hôte.

3.4. Structure et dynamique des génomes

Le génome des bactéries du complexe *B. burgdorferi* s.l. est composé d'un **chromosome** d'environ 910 kilobases (kb), et de **nombreux plasmides** linéaires ou circulaires de tailles variables : de 5 à 60kb environ. Le chromosome contient environ 850 gènes, dont les fonctions sont majoritairement liées aux mécanismes fondamentaux de l'organisme tels que le métabolisme énergétique ou le transport transmembranaire. Les plasmides contiennent quant à eux de nombreux gènes impliqués dans l'adaptation aux variations de milieu dans lequel évolue le pathogène, dans l'hôte ou dans le vecteur. Certains de ces gènes, qui codent pour des protéines de surface sont aussi impliqués dans la pathogénicité. Les plasmides, dont le nombre varie d'un individu à l'autre, sont nommés en fonction de leur structure et de leur taille [Hinnebusch and Barbour, 1992]. Les plasmides circulaires sont ainsi nommés cp et les plasmides linéaires lp. Le plasmide cp26 est donc un plasmide circulaire d'environ 26 kb qui contient entre autre le gène codant pour la protéine de surface C (gène *ospC*). Ce gène est porteur d'une grande diversité génétique en raison de la combinaison de pressions de sélection qui l'influence (sélection disruptives, fréquence-dépendantes négatives) [Brisson and Dykhuizen, 2004 ; Qui *et al.*, 1997 ; Haven *et al.*, 2011]. Il est alors intéressant de rechercher les patrons de diversité spécifique pour voir si l'on peut trouver, le long des différentes unités de réplication qui composent le génome de *Borrelia*, d'autres gènes d'intérêts qui pourraient, par exemple, être impliqués dans la virulence du pathogène.

Au début de ma thèse, il existait **23 souches** de *B. burgdorferi* s.l. dont les génomes ont été entièrement séquencés (Table 1), comprenant quatorze souches de *B. burgdorferi* s.s. [Fraser *et al.*, 1997 ; Schutzer *et al.*, 2011], deux souches de *B. garinii* et deux autres de *B. afzelii* [Casjens *et al.*, 2011b], une souche de *B. bavariensis* [Glöckner *et al.*, 2004], 1 autre

Introduction

de *B. finlandensis* [Casjens *et al.*, 2011a], ainsi que de *B. spielmanii*, *B. valaisiana* et *B. bissetti* [Schutzer *et al.*, 2012].

Table 1 : Description des 23 souches dont les génomes ont été séquencés avant 2014

Souche	Genoespèce	Origine géographique	Source biologique
94a	<i>B. burgdorferi</i> s.s.	USA (New York)	Homme
B31	<i>B. burgdorferi</i> s.s.	USA (New York)	<i>I. scapularis</i>
Bol26	<i>B. burgdorferi</i> s.s.	Italie	<i>I. ricinus</i>
CA-11.2A	<i>B. burgdorferi</i> s.s.	USA (Californie)	<i>I. pacificus</i>
JD1	<i>B. burgdorferi</i> s.s.	USA (Massachusetts)	<i>I. scapularis</i>
N40	<i>B. burgdorferi</i> s.s.	USA (New York)	<i>I. scapularis</i>
WI91-23	<i>B. burgdorferi</i> s.s.	USA (Wisconsin)	Oiseau
ZS7	<i>B. burgdorferi</i> s.s.	Allemagne	<i>I. ricinus</i>
118a	<i>B. burgdorferi</i> s.s.	USA (New York)	Homme
156a	<i>B. burgdorferi</i> s.s.	USA (New York)	Homme
297	<i>B. burgdorferi</i> s.s.	USA (Connecticut)	Homme
29805	<i>B. burgdorferi</i> s.s.	USA (Connecticut)	<i>I. scapularis</i>
64b	<i>B. burgdorferi</i> s.s.	USA (New York)	Homme
72a	<i>B. burgdorferi</i> s.s.	USA (New York)	Homme
Far04	<i>B. garinii</i>	Danemark	Oiseau
PBr	<i>B. garinii</i>	Danemark	Homme
PKo	<i>B. afzelii</i>	Allemagne	Homme
ACA-1	<i>B. afzelii</i>	Suède	Homme
PBi	<i>B. bavariensis</i>	Allemagne	Homme
SV1	<i>B. finlandensis</i>	Finlande	<i>I. ricinus</i>
VS116	<i>B. valaisiana</i>	Suisse	<i>I. ricinus</i>

Souche	Genoespèce	Origine géographique	Source biologique
A14S	<i>B. spielmanii</i>	Pays-Bas	<i>I. ricinus</i>
DN127	<i>B. bissettii</i>	USA (Californie)	<i>I. pacificus</i>

Les premières études du déséquilibre de liaison suggéraient que le complexe d'espèce *B. burgdorferi* s.l. était l'**un des groupes de bactéries les plus clonales** [Maynard Smith and Smith, 1998] et qui ne semblaient donc recombiner que très peu ou pas du tout. Cette conclusion a été appuyée par des études multiloci [Dykhuizen and Baranton, 2001]. Cependant, différentes souches de *B. burgdorferi* s.l. ont été retrouvées en co-infections chez des tiques et des hôtes suggérant que les lignées de *Borrelia* ont des opportunités de recombiner entre elles et des études montrent que des échanges de matériel génétique sont possibles chez *B. burgdorferi* s.l. [Guttman *et al.*, 1996 ; Qui *et al.*, 2002 ; Crowder *et al.*, 2010 ; Tsao *et al.*, 2013]. Ils semblent même être fréquents pour certains loci [Qiu *et al.*, 2004]. Bien que la majorité des études se soient concentrées sur le gène *ospC*, qui présente un fort polymorphisme, et sur la famille des plasmides circulaires cp32, l'observation de transferts génétiques horizontaux a été étendue à tout le génome de *B. burgdorferi* s.s. et le transfert de plasmides entiers a été mis en évidence [Qiu *et al.*, 2004]. Une étude récente basée sur les génomes séquencés (Table 1) suggère que les événements de recombinaison homologue chez *B. burgdorferi* s.l. sont plus fréquents que ce qui avait été initialement imaginé [Haven *et al.*, 2011]. En résumé, bien que *B. burgdorferi* s.l. ne soit pas strictement clonal, les événements de recombinaison apparaissent comme rares en général et sont presque inexistantes sur le chromosome mais largement plus fréquents sur les plasmides [Qiu *et al.*, 2004] et se concentrent au niveau de points-chauds tel que le gène *ospC* où ils pourraient jouer un rôle important dans la création de diversité génétique [Barbour and Travinsky, 2010]. Cependant, il serait intéressant d'acquérir des informations à partir d'un échantillonnage local et aléatoire et de comparer ces résultats avec ceux disponibles à partir des souches de *B. burgdorferi* s.s. de la Table 1. De plus, les taux de recombinaison d'autres génoespèces que *B. burgdorferi* s.s. pourraient être estimés afin d'avoir une idée des différences qui peuvent exister entre les différentes espèces du complexe.

4. Problématiques et plan de thèse

Les données actuelles indiquent que les différentes génoespèces du genre *Borrelia*

Introduction

interagissent avec différents vecteurs et différents hôtes. Dans cet environnement complexe, la question principale qui se pose est : « Comment circulent les agents pathogènes du genre *Borrelia* entre leurs hôtes et leurs vecteurs ? ». D'un point de vue épidémiologique et pour améliorer la prévention dans les zones endémiques, il s'agit de comprendre d'où proviennent les **sources d'infection**, de définir les différents **cycles de transmission** de la bactérie et de déterminer lesquels sont à l'origine du **risque pour l'homme**. Pour cela, il est tout d'abord nécessaire d'être capable de définir les différentes espèces et les différents génotypes de *Borrelia* portés par les tiques et les hôtes, pour ensuite les étudier et caractériser les processus mis en jeu dans leur évolution.

Le **premier chapitre** de ma thèse aborde le potentiel des approches de génétique des populations et de génomique comparative pour comprendre les processus de diversification du complexe d'espèces *B. burgdorferi* s.l.. Pour cela, les frontières entre les différentes génoespèces du complexe ont été étudiées à l'échelle de différentes populations de *B. burgdorferi* s.l.. Les génomes partiels de 63 isolats, appartenant à trois génoespèces du complexe et échantillonnés en Alsace, une région où la maladie de Lyme est endémique, ont été obtenus par séquençage haut-débit. Des analyses phylogénétiques et de modélisation des généalogies, ont notamment permis de caractériser l'isolement génétique entre les différentes génoespèces et d'étudier la distribution des transferts génétiques inter- et intra-spécifiques. Des statistiques descriptives couramment utilisées en génétique des populations et une approche de modélisation phylodynamique nous ont permis d'étudier les différences de structure de populations pour les différentes génoespèces. Pour finir, cette approche nous a également permis de détecter des régions du génome pouvant être soumises à des pressions de sélection particulières et de souligner la particularité du gène *ospC* en raison des pressions de sélection auxquelles il est soumis.

Dans le **second chapitre** de la thèse, il a été question de décrire les associations entre les génotypes des espèces de *B. burgdorferi* s.l. et les différentes espèces d'hôtes échantillonnées. Dans ce cadre, nous avons analysé les séquences haut-débit d'un gène de ménage (*rplB*) et du gène *ospC* qui ont des fonctions différenciées et qui sont soumis *a priori* à des pressions de sélection différentes. Ces données ont en effet permis d'étudier la diversité de génotypes de *Borrelia* portée par environ 230 hôtes ont rendu possible l'étude des co-infections des hôtes par différents génotypes de *Borrelia*.

Le **troisième chapitre** de la thèse aborde les modalités de circulation des *Borrelia* à l'échelle des communautés d'hôtes et de tiques. Les données de séquences des gènes *rplB* et *ospC* des *Borrelia* infectant les hôtes vertébrés échantillonnés ont été complétées par celles des *Borrelia* portées par 420 tiques vectrices. Les contributions des différents hôtes à la circulation des bactéries chez les tiques ont été estimées grâce i) à la comparaison de la diversité de séquences chez les hôtes vertébrés échantillonnés avec celle observée chez les tiques vectrices et ii) à la construction d'un nouveau modèle statistique adapté aux données en notre possession.

Dans un **dernier chapitre**, les résultats obtenus à l'aide des données acquises seront synthétisés, discutés et mis en relation avec différentes perspectives de recherches.

CHAPITRE 1

GENOMIQUE COMPARATIVE DU COMPLEXE D'ESPECES *B. BURGDORFERI* S.L.

Résumé de l'article 1

Comme nous avons pu le voir, dans le cadre de l'étude des maladies zoonotiques à transmission vectorielle, un défi important est de comprendre la diversification et les modalités de circulation des agents pathogènes entre leurs hôtes et leurs vecteurs. Ce chapitre s'articule autour de l'étude du potentiel des approches de génomique comparative et génomique des populations pour comprendre ces processus avec, pour système modèle, le complexe d'espèces *B. burgdorferi* s.l.. Pour cela, 63 souches, appartenant à trois génoespèces du complexe, ont été isolées à partir de tiques à l'affût échantillonnées en Alsace, une région où la maladie de Lyme est endémique, par des membres de l'Institut Pasteur (Centre National de Référence *Borrelia*). Leurs génomes ont ensuite été séquencés. J'ai effectué le mapping des séquences obtenues sur des génomes de référence et construit les alignements des différentes unités de réplication. C'est à partir de ces alignements que j'ai réalisé, avec l'aide de nombreux membres de l'unité d'Epidémiologie Animale, les différentes analyses dont les principaux résultats sont présentés ci-dessous. Ce travail a été publié dans la revue *PLoS ONE* en Avril 2014.

Distribution de la diversité génétique au sein du complexe

D'une part, les analyses phylogénétiques conduites sur les séquences chromosomiques et les résultats de modélisation des généalogies, à l'aide de coalescents, ont révélé un degré d'isolement génétique très important entre les génoespèces échantillonnées *B. burgdorferi* s.s., *B. garinii* et *B. afzelii*. Le taux de recombinaison au sein de ces espèces a été estimé être environ 50 fois plus élevé que le taux de recombinaison entre les espèces. De plus, les relations phylogénétiques observées entre les séquences de nos souches et celles de références, suggèrent que la diversification des lignées de *B. burgdorferi* s.s. et de *B. garinii* échantillonnées a largement précédé la mise en place des populations dans le site d'échantillonnage.

D'autre part, les résultats concernant la divergence moyenne observée i) entre les souches de *B. burgdorferi* s.s. et la souche de référence de *B. filandensis* SV1 ainsi ii) qu'entre les souches de *B. garinii* et la souche de *B. bavariensis* PBi, suggèrent que SV1 et PBi n'ont pas encore totalement divergé des lignées desquelles elles sont proches, ce qui pourrait être à l'origine d'une différenciation incomplète des lignées. Ces résultats posent notamment question concernant la quantité d'information qu'il est nécessaire d'utiliser pour délimiter différentes génoespèces.

Chapitre 1

Dynamiques des populations hôtes-dépendantes

Les statistiques descriptives couramment utilisées en génétique des populations et un modèle phylodynamique que nous avons construit, nous ont permis d'apporter des éléments pour expliquer les patrons de diversité différents observés pour les espèces *B. garinii* et *B. burgdorferi* s.s.. En effet, les résultats fournis suggèrent que la taille de population et que le taux de migration entre les deux sites d'échantillonnage, étaient plus élevés dans les populations de *B. garinii* que dans les populations de *B. burgdorferi* s.s.. De plus, les taux de renouvellement estimés pour les hôtes infectés par les deux espèces de *Borrelia*, sont différents et plus importants pour *B. burgdorferi* s.s.. Ces résultats fournissent la preuve que les différents spectres d'hôtes des lignées de *B. burgdorferi* s.s. (principalement associées aux petits mammifères) et de *B. garinii* (normalement associées aux oiseaux) conduisent à des dynamiques de populations aux profils dissemblables.

Pressions de sélection et diversification du gène lié à l'infection, ospC

Les patrons des statistiques descriptives ont permis de souligner la singularité de la région du plasmide cp26 proche du gène *ospC* qui est soumis à des pressions de sélection diversifiantes et fréquence-dépendantes négatives. Les événements de recombinaison fréquents impactant le gène et ses régions flanquantes, sont la preuve de co-infections passées entre *B. burgdorferi* s.s., *B. afzelii*, et *B. spielmanii* suggérant que ces trois génoespèces circulent chez les mêmes hôtes pour le site étudié en Alsace. Dans le cadre de l'étude des cycles de transmission de la bactérie entre les différents hôtes qu'elle peut infecter, le gène *ospC* semble un bon marqueur fonctionnel et souligne l'importance de prendre en compte les co-infections.

ARTICLE 1

Comparative population genomics of the *Borrelia burgdorferi* species complex reveals high degree of genetic isolation among species and underscores benefits and constraints to studying intra-specific epidemiological processes

Maude Jacquot, Mathieu Gonnet, Elisabeth Ferquel, David Abrial, Alexandre Claude, Patrick Gasqui, Valérie Choumet, Myriam Charras-Garrido, Martine Garnier, Benjamin Faure, Natacha Sertour, Nelly Dorr, Jocelyn De Goër, Gwenaël Vourc'h, Xavier Bailly



Comparative Population Genomics of the *Borrelia burgdorferi* Species Complex Reveals High Degree of Genetic Isolation among Species and Underscores Benefits and Constraints to Studying Intra-Specific Epidemiological Processes

Maude Jacquot^{1*}, Mathieu Gonnet¹, Elisabeth Ferquel^{2a}, David Abrial¹, Alexandre Claude^{1ab}, Patrick Gasqui¹, Valérie Choumet^{2ac}, Myriam Charras-Garrido¹, Martine Garnier^{2ad}, Benjamin Faure^{1ae}, Natacha Sertour^{2af}, Nelly Dorr¹, Jocelyn De Goër¹, Gwenaél Vourc'h¹, Xavier Bailly¹

¹ INRA, UR346 Épidémiologie Animale, Saint Genès Champanelle, France, ² Institut Pasteur, CNR *Borrelia*, Paris, France

Abstract

Lyme borreliosis, one of the most frequently contracted zoonotic diseases in the Northern Hemisphere, is caused by bacteria belonging to different genetic groups within the *Borrelia burgdorferi* species complex, which are transmitted by ticks among various wildlife reservoirs, such as small mammals and birds. These features make the *Borrelia burgdorferi* species complex an attractive biological model that can be used to study the diversification and the epidemiology of endemic bacterial pathogens. We investigated the potential of population genomic approaches to study these processes. Sixty-three strains belonging to three species within the *Borrelia burgdorferi* complex were isolated from questing ticks in Alsace (France), a region where Lyme disease is highly endemic. We first aimed to characterize the degree of genetic isolation among the species sampled. Phylogenetic and coalescent-based analyses revealed clear delineations: there was a ~50 fold difference between intra-specific and inter-specific recombination rates. We then investigated whether the population genomic data contained information of epidemiological relevance. In phylogenies inferred using most of the genome, conspecific strains did not cluster in clades. These results raise questions about the relevance of different strategies when investigating pathogen epidemiology. For instance, here, both classical analytic approaches and phylodynamic simulations suggested that population sizes and migration rates were higher in *B. garinii* populations, which are normally associated with birds, than in *B. burgdorferi* s.s. populations. The phylogenetic analyses of the infection-related *ospC* gene and its flanking region provided additional support for this finding. Traces of recombination among the *B. burgdorferi* s.s. lineages and lineages associated with small mammals were found, suggesting that they shared the same hosts. Altogether, these results provide baseline evidence that can be used to formulate hypotheses regarding the host range of *B. burgdorferi* lineages based on population genomic data.

Citation: Jacquot M, Gonnet M, Ferquel E, Abrial D, Claude A, et al. (2014) Comparative Population Genomics of the *Borrelia burgdorferi* Species Complex Reveals High Degree of Genetic Isolation among Species and Underscores Benefits and Constraints to Studying Intra-Specific Epidemiological Processes. PLoS ONE 9(4): e94384. doi:10.1371/journal.pone.0094384

Editor: Brian Stevenson, University of Kentucky College of Medicine, United States of America

Received: December 10, 2013; **Accepted:** March 13, 2014; **Published:** April 10, 2014

Copyright: © 2014 Jacquot et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by funding from the French National Institute for Agricultural Research (INRA) through the 'AIP Bioressources 2009' initiative, the Auvergne region through the 'Lifegrid' program, and the European Union through the European Regional Development Fund (ERDF). Maude Jacquot received PhD funding from the French Ministry of Higher Education and Research (MESRI). Mathieu Gonnet was supported by the ERDF and by the OSCAR project of the French National Research Agency. Alexandre Claude was supported by the 'Lifegrid' program. Interactions between members of the INRA and the Pasteur Institute are promoted by the 'Tick and Tick-Borne Diseases' (TMT) group, which is part of the Research Consortium on Biological Interactions (GdR REID). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: maude.jacquot@gmail.com

^a Current address: Institut Pasteur, Département de Virologie, Paris, France

^b Current address: CNRS IN2P3, Laboratoire de Physique Corpusculaire, Aubière, France

^c Current address: Institut Pasteur, Unité des Interactions Moléculaires Flavivirus-Hôtes, Paris, France

^d Current address: Institut Pasteur, Unité de Recherche et d'Expertise des Bactéries Pathogènes Entériques, Paris, France

^e Current address: INRA, Centre de Traitement de l'Information Génétique, Jouy-en-Josas, France

^f Current address: Institut Pasteur, Unité Biologie et Pathogénicité Fongiques, Paris, France

Introduction

Zoonotic diseases caused by pathogens that are transmitted among different host species represent an emergent threat for human health [1]. However, the study of these pathosystems is

hampered by their complexity, as each pathogen may have multiple potential reservoirs. In vector-borne systems, in which infected vectors feed on and transmit the pathogen to several hosts, it is possible to characterize pathogen diversity using population genomic studies. These studies, which use information obtained

from the genome of pathogens isolated from questing vectors, offer researchers the opportunity to indirectly study the spread of pathogens within and among host communities. However, genetic information can be affected by many processes, and untangling various lines of evidence in order to obtain a coherent picture of the evolutionary history of a population represents a significant challenge for population genomic studies [2].

In particular, the selective constraints on pathogen genomes can be highly heterogeneous. For example, infection-related genes are expected to be affected by disruptive and/or negative frequency-dependent selection due to the molecular interactions occurring among pathogens, hosts, and the hosts' immune systems. Therefore, within bacterial lineages associated with similar hosts these genes would be predicted to share similar features due to host-driven selective sweeps, DNA exchange by recombination, or convergent evolution. Patterns of diversity in these genes could thus potentially reveal which lineages infect overlapping host communities. Alternatively, housekeeping genes, which are involved in basic cell cycle and metabolic functions, are mostly subject to purifying selective pressures [3]. They are more likely to contain relevant information about gene flow among pathogen populations and other demographic events, especially if pathogen lineages have had the opportunity to recombine [4]. Indeed, most regions of the genome should not be affected by host driven selection if the recombination rate is sufficiently high [5], which is an advantage for researchers examining demographic and/or epidemiological processes. Infection-related genes and other genome regions can thus provide complementary information on bacterial transmission. One of the main challenges of population genomic approaches is to integrate hypotheses about patterns of diversity that are observed at the scale of individual genes or genome regions into evolutionary scenarios that are coherent at the level of the whole genome.

Using appropriate genome regions, analyses of demographic and epidemiological processes from population genomic data rely on models that describe both the way pathogens spread within and among susceptible individuals, as well as processes that govern genome evolution. These two aspects have historically been investigated independently in the literature. On the one hand, the spread of pathogens has often been studied by epidemiologists using population dynamic models that take into account complex infection and transmission processes [6], but most of the time neglect the diversity of pathogens. On the other hand, the evolution of genomes has been studied using tools of population genetics that often assume basic demographic hypotheses, as in the Wright-Fisher [7] or the Moran [8] model. The increasing prevalence of molecular epidemiology studies highlights the need to close the gap between these two approaches, a feature that defines the burgeoning field of phylodynamics. While phylodynamic approaches have produced promising results in the study of epidemics of fast-evolving viruses, their applicability to endemic, slow-evolving, bacterial pathogens remains to be assessed [9]. A particular challenge for the use of phylodynamic models is that genomic polymorphisms represent the footprints of processes that have occurred across multiple geographic and/or time scales. For example, phylogeographical studies have shown that a population's history plays a major role in shaping its current diversity patterns. Large-scale patterns, such as the occurrence of major clades within a species can be due to ancestral differentiation in independent geographic locations (called refugia), while the distribution of diversity on a smaller scale is subsequently shaped by more recent colonization and migration events [10]. The picture is even more complex in multi-host pathogen systems, in which major clades can emerge either from isolated geographic

regions where the pathogen circulates or from independent reservoir hosts as defined by Haydon *et al.* [11]. Then, the diversity that is maintained in a given reservoir at endemic equilibrium would depend not only on epidemiological parameters such as the number of infected hosts and transmission parameters, but also on the processes that generate sequence diversity in that population [12].

In core regions of bacterial genomes, sequence diversity emerges through both mutation and homologous recombination. Twenty years ago, Maynard Smith *et al.* revealed the impact of recombination on the evolution of bacterial pathogens, which can be hidden by differences in the reproductive success of genotypes [13]. In the so-called epidemic population structure, the uneven frequencies of recombinant lineages lead to high statistical associations among genotypes at different loci, i.e. linkage disequilibrium. By selecting a subsample of unique multilocus genotypes within a population, Maynard Smith and colleagues were able to more easily identify statistical evidence for recombination. Although still relevant, this sub-sampling approach has major drawbacks, as it is difficult to quantitatively study the processes that shape the distribution of diversity in subsamples that have been defined by genetic criteria *a priori* [14]. Therefore, studies of pathogen diversity commonly use hierarchical sampling. Factors that structure diversity, such as bacterial taxonomy or sampling location, are defined *a priori*, and a random sample of strains is obtained for each selected factor [15]. Then, the accuracy of species delineations (or whichever structuring factor is used) can be evaluated *a posteriori* with regard to the observed patterns of polymorphism before fitting more complex evolutionary and epidemic models to the genomic data.

To assess the potential of population genomic approaches in the study of multi-host zoonotic disease, we chose the *Borrelia burgdorferi* species complex as our model system. This species complex includes the bacteria that cause Lyme borreliosis, one of the most common vector-borne diseases in the Northern Hemisphere. In Europe, most human cases are caused by *Borrelia afzelii*, *Borrelia burgdorferi sensu stricto* (s.s.), and *Borrelia garinii* [16], which are transmitted by the tick *Ixodes ricinus* among different reservoir species. The ticks can become infected during their first blood meal, which occurs before they molt from larvae into nymphs. They can also acquire and/or transmit bacteria during their other blood meal(s), which occur before the molt from nymph to adult (all ticks) and before laying eggs (females only).

Bacteria within the *B. burgdorferi* complex do not share a common, fixed set of host species (referred to hereafter as host range) [17]. Phylogenetic analysis of multi-locus sequence typing (MLST) data suggests that similar host ranges have evolved several times within this complex, so that bacteria associated with similar hosts are not clustered in clades, but rather, dispersed throughout the whole phylogeny [18]. *Borrelia garinii* and *Borrelia valaisiana*, which are only distantly related to each other, both infect birds [19–22], whereas *B. afzelii*, *Borrelia spielmanii*, and *Borrelia bavariensis*, which do not cluster together in phylogenies, infect small mammals [23–25]. Furthermore, the host range of *B. burgdorferi* s.s. appears to be broader than that of most *Borrelia*, as this species has been identified in both small mammals and birds [26–28]. However, despite the large number of host shifts that have occurred through the course of the evolution of this species complex, the influence of mutation and recombination events on these bacteria's ability to adapt to new hosts is poorly understood [29].

Different approaches have been used to investigate the influence of host communities on the diversity of the *B. burgdorferi* species complex. From a population genetic point of view, differences in

the population size and the migration potential of reservoir species have been shown to influence MLST diversity patterns of these bacterial species at a continental scale [30], but there is a lack of information at smaller geographical scales. From an epidemiological point of view, a statistical model has been used to estimate the respective contributions of different host species to the infection of ticks by genotypes of *B. burgdorferi* s.s. in the USA [31]. This approach takes advantage of the associations found among genotypes of the *B. burgdorferi* species complex obtained from different populations and host species at the infection-related gene *ospC* [32,33].

The development of genetic and genomic tools has allowed the in-depth investigation of genes involved in infection phenotypes within the *B. burgdorferi* species complex [34]. A recent study of synonymous and non synonymous mutation rates from multiple genomes has highlighted a potential role for lipoproteins-encoding genes in the adaptation of the *B. burgdorferi* species complex to hosts [35]. The same sample of genomes, which was predominantly obtained from strains of *B. burgdorferi* s.s., was also used to investigate evolutionary processes within this species [36], although inter-specific patterns were described only superficially. While the focus of this study was to investigate recombination patterns and selective constraints within *B. burgdorferi* s.s., the selected isolates were chosen in order to capture a maximum of genetic diversity [37]. However, as described above, this type of sampling could reveal a different genetic structure compared to a random regional sampling.

Here, we investigated the genomic diversity of strains of *B. burgdorferi* s.s., *B. garinii*, and *B. afzelii* isolated from questing ticks in France, in a region where Lyme borreliosis is highly endemic. Using a hierarchical sampling scheme at a regional scale, we aimed to investigate the genetic diversity within and isolation among species of the complex. We also evaluated multiple models, involving both evolutionary and epidemiological constraints, and examined the information obtained from them to understand the forces that have shaped diversity within this species complex. This lead to different hypotheses regarding the evolution and the epidemiology of these bacteria. The data gathered here are then discussed in the context of what is currently known about the ecology and evolution of these vector-borne pathogens.

Materials and Methods

Ethics statement

Isolates of the *B. burgdorferi* species complex were recovered from questing nymphs and adults of *I. ricinus* that were sampled at two sites largely covered by dense and continuous forests; these sites were located near the towns of Munster and Guebwiller, in Alsace (France) [38] (Figure 1). No specific permission is required by French law to sample ticks and to perform field studies in these locations. Our study did not involve any endangered or protected species.

Bacterial strains and sequencing procedures

Ticks were collected at each site in two consecutive years (2003 and 2004) by dragging a large piece of cotton fabric across the vegetation and leaf litter [38]. Ticks were then kept alive in individual tubes until they could be processed. Bacterial isolation was performed by incubating ticks individually at 32°C in 5 ml of BSK-H medium (Sigma-Aldrich, St. Louis, USA) without antibiotics for 8 weeks. Bacterial growth was checked weekly via dark field microscopy. Cultures that showed evidence of spirochete growth were maintained until they reached high densities. After that PCR-RFLP was used to assign each culture to a *Borrelia*

species [39]. Lastly, each culture was supplemented with 60% glycerol in an equal amount to that of the culture and kept at -80°C, pending further analyses.

From the different isolates, we chose 63 strains, to be sequenced. As we aimed i) to study species delineations within the *B. burgdorferi* complex, ii) to examine the distribution of diversity within sympatric species of the complex at a regional scale, and iii) to compare the genetic structures of different bacterial species, we selected a set of strains that included 25 genotypes of *B. burgdorferi* s.s., 34 genotypes of *B. garinii*, and 4 genotypes of *B. afzelii*. Within each species, genotypes were randomly selected. The size of the three samples was determined by the rate of success of the isolation procedure rather than by the observed frequency of the different species within ticks. Moreover, we selected isolates with the goal of having a similar number of strains for each sampling site within each species. We also tried to select strains that were sampled during the same year in order to limit confounding factors: all *B. garinii* and *B. afzelii* strains were isolated in 2004, whereas an equal number of *B. burgdorferi* s.s. strains were isolated in 2003 and in 2004 (Table S1).

To obtain enough material for sequencing, bacteria were iteratively cultured using 50 ml of BSK-H medium (Sigma-Aldrich) until high bacterial densities were reached. All strains underwent fewer than 15 passages from tick incubation to DNA extraction in order to limit the possibility of plasmid loss. Independent DNA extractions were performed using the DNeasy Blood & Tissue Kit (Qiagen, Venlo, Netherlands). The quantity of DNA after extraction was measured by spectrophotometry at 260 and 280 nm (Nanodrop, Thermo Fisher Scientific, Waltham, USA), and DNA extractions were sent to Genoscreen (Lille, France) to be tagged with standard multiplex identifiers (MID, Roche, Basel, Switzerland). Samples were then mixed to prepare libraries, which were distributed among three-quarters of a GS FLX Titanium (Roche) sequencing plate.

In order to analyze sequence data, raw reads were first mapped on reference sequences. Contigs were aligned with reference genomes. The genetic structure of the sample was studied using single nucleotide polymorphisms (SNPs)-based analyses at intra- and inter-specific levels. Additionally, phylogenetic analyses were performed at the same scales. Loci showing atypical polymorphism patterns were identified using allelic spectrum-based approaches and the functions they encode were analyzed. Finally, the results of SNP-based analyses on the chromosome, which revealed homogeneous polymorphism patterns, were used to fit a long term coalescent based model and a basic, shorter term, phylodynamic model. These different steps are described in details below.

Reconstruction of genetic sequences

To obtain robust data, each sequence read was mapped independently onto reference sequences of *B. burgdorferi* s.s., *B. bavariensis*/*B. garinii*, and *B. afzelii*. For each of the three mappings, reference sequences of the chromosome, the circular plasmid cp26, and the linear plasmid lp54 were chosen from public databases (Table S2). We focused on the chromosome and the plasmids cp26 and lp54 because the gene content and the synteny of these replication units are relatively conserved among strains of the *B. burgdorferi* species complex [35]. Other replication units tend to show more polymorphism, both in terms of gene copy number and in term of gene arrangement. We did not include them in this study because of the problems they raised with sequence assembly and the specific analyses they required. To ensure accurate mapping in divergent genomic regions while maintaining sufficient stringency for subsequent analyses, we used GSMAPPER software (Roche) parameterized as follows: the length of seeds used to

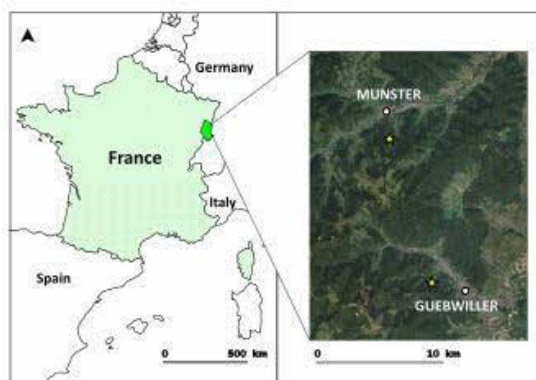


Figure 1. Map of the sampling sites. Tick sampling sites (yellow stars) were located near the towns of Munster and Guebwiller in northeastern France. The data used to construct the map were obtained from BD-Ortho in 2011 (IGN; National Institute of Geographic and Forest Information, Paris, France). doi:10.1371/journal.pone.0094384.g001

anchor alignments was fixed to 10 base pairs (bp) and the identification of three seeds per sequence was required for alignment analysis. Identity thresholds were fixed at 60 bp and 60% identity. The results of mapping onto a given replicon of a given species were stored if more than 90% of the reference sequence was covered. Then, a strict consensus sequence was generated for each replication unit of each bacterial isolate from the alignment of the stored mapping results. This was performed using a program we developed in Pascal; unless otherwise indicated, programs developed in Pascal were used for all analyses described below.

Finally, to compare our genomic data to published sequences (Table S3), we constructed multiple alignments for each of the three studied replication units. PROGRESSIVEMAUVE software [40] was used to define homologous sequence regions using default parameters, and local alignments were refined using MUSCLE software [41] on contiguous windows of 1 kilobase pairs (kb) in length.

Single-nucleotide polymorphism analyses

SNPs were identified using multiple alignments. To explore the population structure of *B. burgdorferi* s.s. and *B. garinii*, a set of SNPs was for each species that included all polymorphic sites in the chromosome that were identified in at least 90% of strains. The two sets of SNPs were analyzed separately using the program STRUCTURE v2.3.4 to identify potential populations and explore their degree of admixture [42]. This method has been developed to estimate allele frequencies in an user-defined number of populations and to assign individuals to these populations based on a Markov Chain Monte Carlo (MCMC) scheme. For each species, we performed analyses assuming correlations among linked loci and allowing admixture among potential populations. We investigated models with K, the number of populations, ranging from 1 to 15. For each value of K, five different runs were performed in which the MCMC algorithm was run for a 25,000-iteration burn-in step and followed by 25,000 further iterations. The most appropriate K values were chosen by taking into account the observed likelihood and the reproducibility of results.

In order to limit the impact of base-calling errors (which should be independent among individuals) and to obtain sufficient

statistical power for subsequent analyses, we selected SNPs that: i) were present in all individual strains and ii) had at least two alleles with frequencies higher than 10%. Differentiation measures reflect the genetic variability among groups of individuals relative to the variability of a whole sample [43]. In order to explore the distribution of genetic variability, we chose to measure differentiation, using H_{ST} values [44]. H_{ST} is a multiallelic numerical analog of Wright's F_{ST} [45], and is obtained using the formula $H_{ST} = 1 - H_S / H_T$, where H_S and H_T represent Nei diversity indices within and among populations, respectively. We calculated H_{ST} for selected SNPs at different levels of genetic resolution: i) among species of the complex, to assess their genetic isolation from each other and ultimately identify evidence of inter-specific homologous recombination; and ii) within studied species and between sampling sites, in order to investigate geographic isolation and identify genomic regions evolving under particular selective pressures [46]. As no evidence of genetic differentiation was observed between strains of *B. burgdorferi* s.s. sampled in 2003 and those sampled in 2004 (data not shown), we did not consider this potential structuring factor in the analyses described above. The significance of the H_{ST} values was tested with a Monte-Carlo approach ($p = 0.05$): for each SNP, 1000 simulated datasets that distributed individual genotypes at random among populations were created. Additionally, we compared the distribution of H_{ST} values in bacterial populations of *B. burgdorferi* s.s. and *B. garinii* that were sampled in different locations using a Wilcoxon rank-sum test with continuity correction, performed in R [47].

As the distribution of H_{ST} values depends on the genetic linkage between SNPs, we investigated the impact of homologous recombination on the genetic diversity of the studied strains. Standardized linkage disequilibrium measures between pairs of SNPs were obtained using the D' statistic [48], and the approach proposed by Hedrick [49] to summarize linkage disequilibrium among different allelic combinations. D' values were computed for each pair of SNPs within the following groups: i) all strains, ii) all strains assigned to the same species, and iii) conspecific strains isolated from both of our sampling sites in Alsace. For each species, we then explored the relationship between D' values and physical distance.

Phylogeny-based analyses

To conduct detailed analyses of the phylogenetic relationships among genotypes using the alignments of the three replication units, contiguous 1-kb-long windows were created. We decided not to work at the gene level in order to standardize the amount of phylogenetic information contained in each unit. A phylogenetic search that applied a maximum-likelihood approach using PhyML software [50], was performed on each window. The most appropriate model of evolution was chosen for each alignment based on the Akaike Information Criterion (AIC) [51] using the APE library in R [52]. For each window, the maximum-likelihood tree was screened to assess whether different species or groups of species were monophyletic. This analysis provided information about which genome regions supported the consensus phylogeny and, conversely, which genome regions were potentially affected by interspecific recombination or other evolutionary events.

To illustrate global genetic relationships within and among species, phylogenetic networks, based on chromosomal sequences of sampled genotypes and reference genomes, were constructed with SPLITSTREE 4 software [53] using the Neighbor-Net method [54]. To generate these networks, we first created an alignment of chromosome sequences, and then used this alignment to compute a distance matrix in PAUP* 4.0 b10 [55] using a GTR+I+G model [56,57] with the following settings: the substitution rate matrix was

estimated via maximum likelihood assuming empirical nucleotide frequencies, while the proportion of invariable sites and the shape parameter of the gamma distribution were fixed at the respective mean values that were obtained from the maximum likelihood phylogenetic analyses of the 1-kb-long windows described above.

Next, we assessed whether the percentage of windows in which a given species was not monophyletic could be explained not by recombination but simply by a lack of resolution. With this aim, we simulated 300 alignments of 63 sequences of 1000 kb according to i) the GTR+I+G model described above and ii) a neighbor-joining tree [58] obtained from the average distance matrix described in the previous paragraph and based on the appropriate model of sequence evolution that described divergence patterns among the studied genomes. For each simulated alignment, 1,000 1-kb-long contiguous windows were delineated and a maximum likelihood phylogeny based on the assumed model of sequence evolution was obtained for each window. For each phylogeny, the monophyly of species was assessed as described for observed data. Afterwards, the distribution of the percentage of windows in which species were not monophyletic within the simulated alignments was compared to that obtained from observed data. Simulations and phylogenies were generated using the Bio++ C++ library [59].

Further alignment-based analyses

To describe diversity within each of the three species, two statistics were computed using contiguous 1-kb windows of alignments of each replicon for each species: i) Watterson's θ_s , which is based on the number of segregating sites along the alignment [60], and ii) Tajima's θ_s , which is based on the average divergence among samples [61]. Furthermore, to obtain insight into deviation from demographic equilibrium and selective neutrality within the *B. burgdorferi* species complex, values of Tajima's D [62] were calculated for the 1-kb windows within each replication unit for each species. Again, we chose not to work at the gene level to standardize the amount of available information among analyses; annotated genes in genomes of the *B. burgdorferi* species complex vary widely in length, with some too short to provide an adequate amount of mutation for this type of analysis. Gapped sites were not considered, as in Tajima's original publication. Tajima's D statistic measures the difference between the two estimators of θ previously described, standardized by the variance of this difference. Under the neutral evolutionary model, which assumes that polymorphisms segregate at mutation-drift equilibrium, Tajima's D is expected to be null. Positive Tajima's D value can be induced by balancing selection, due to the maintenance of highly divergent variants. Conversely, negative Tajima's D values can be the result of purifying selection or a selective sweep, which both result in an excess of weakly divergent alleles. However, these two patterns (positive or negative) may alternatively be due to demographical/epidemiological processes that affect the whole genome. Trying to distinguish between potential whole-genome versus gene-region-specific process, we looked for windows within each species' sequence that were characterized by a highly structured pattern of diversity. These genome regions might be evolving under host-driven selective pressures. Towards this end, we first obtained the distribution containing Tajima's D values for every window of chromosomal data for each studied species. As Tajima's D values were, on average, negative (see results), standard tests that identify deviations from neutrality did not provide relevant information. We thus studied chromosomal Tajima's D values and defined a threshold that distinguished relatively high measures within each species, based on the 95th percentile of the distribution of Tajima's D values within that species. Then, we identified and selected the

windows in the three replication units that had Tajima's D values that were higher than the defined thresholds.

Genes that overlapped the selected windows were identified and compared among the different species, and we created a non-redundant list containing the selected genes. The function and the cellular localization of the proteins encoded by the selected genes were studied using the online pipeline SLEP [63]. The distribution of genes among the different functional categories was compared to the results obtained from the total proteome of the three studied replication units.

Differentiation measures, Tajima's D values, and linkage disequilibrium patterns revealed that a 4-kb region around the *ospC* gene on the cp26 plasmid was characterized by a peculiar polymorphism pattern. To illustrate the amount of incongruent phylogenetic information in *ospC* sequences and flanking regions, Neighbor-Net networks, based on uncorrected p-distances, were obtained from alignments of the *ospC* gene, the 2,000 bp before and after the gene, and regions located further upstream and downstream.

Long-term coalescent based model

As described in the introduction, genome sequences contain evidences of processes that occur at different time scale. In order to study the long-term processes that have shaped the diversity of the *B. burgdorferi* species complex, we developed a coalescent model to investigate the divergence between *B. burgdorferi* s.s. and *B. garinii*. This model used species-specific properties to simulate sequence datasets and employed Approximate Bayesian Computations (ABC) [64] to obtain inferences regarding parameters of interest.

The original coalescent model can be defined as a Markov process describing how, assuming a Wright-Fisher or Moran model of genetic variation, sampled individuals would share a common ancestor (coalesce) as one goes back in time [65]. A central result of this is that the number of generations that one must go back in time to the next common ancestor shared by one pair of samples in a population, i.e. a coalescence event, follows an exponential distribution, the shape of which depends on both the number of samples included in the analysis and the effective population size. Using an iterative procedure, the properties of an entire genealogy of samples can be recovered and the genealogy can then be used to simulate sequence data.

In its simplest form, the coalescent model assumes a lack of recombination and a lack of population structure; however the model has been extended to allow for such evolutionary events as homologous recombination [66] or population subdivision [67]. For example, to include recombination in the model, the exponential distribution is modified so that it describes the time before either a coalescence event or a recombination event. Similarly, to take population subdivision into account population, the exponential distribution is modified so that: i) it describes the time interval to previous coalescence event in each population, and ii) the population size of the two populations can differ, using scaling factors.

The model we developed for this study assumes two steps. During the first step, two populations, representing *B. burgdorferi* and *B. garinii*, are simulated, characterized by constant effective population sizes of $c_1 \cdot N$ and $c_2 \cdot N$, respectively. Coalescence events are only possible within populations, and unidirectional homologous recombination is allowed both within and between populations. In each generation, for each individual, any sequence site in a given population is susceptible to intra-population homologous recombination at respective rates of r_{intra1} and r_{intra2} . Similarly, recombination events can occur between individuals belonging to different populations at rates r_{inter1} and r_{inter2} . After $M \cdot (c_1 \cdot N + c_2 \cdot N)$

generations during which the two populations evolve independently, a new step is initiated. The two populations merge into a single ancestral population, with effective population size $c_3 \cdot N$ and intra-population recombination rate $r_{\text{intra}3}$. As it is impossible for the populations to coalesce into a single common ancestor if the probability of recombination is higher than the probability of coalescence, if the common ancestor has not been reached above $(M+S) \cdot (c_1 \cdot N + c_2 \cdot N)$ generations, recombination is stopped. Once a complete genealogy is obtained, mutations are added along branches at a rate μ per site per generation according to the Jukes-Cantor model [68].

Various nested models were used to simulate alignments of 10,000-bp sequences that included 23 samples for the first population (*B. burgdorferi* s.s.) and 32 samples for the second (*B. garinii*). In our initial model (M0), all the variables described above were assumed to be independent. In model M1 and all subsequent models, we assumed $r_{\text{intra}1} = r_{\text{intra}2}$. In models M2 and M3, we assumed, respectively, $c_1 = c_2 = c_3$ and $r_{\text{intra}1} = r_{\text{intra}2} = r_{\text{intra}3}$, while the M4 model incorporated both of these assumptions. We finally investigated an M5 model that assumed $r_{\text{intra}1} = r_{\text{intra}2} = r_{\text{intra}3} = r_{\text{intra}4}$ and $c_1 = c_2 = c_3$. After exploring model behavior, we performed 50,000 simulations for each model, fixing the parameters as follows to shorten computation time: $N = 1e+09$, $\mu = 1e-07$ and $S = 5$. Based on the relationship we observed between linkage disequilibrium and physical distance on the genome (see results), we assumed a fixed recombination fragment length of 500 bp, a length that is consistent with current hypotheses about recombination track lengths [34]. For each simulation, we sampled values for the different parameters of interest in uniform distributions ranging from 0 to 1e-08 for $r_{\text{intra}1}$, from 0 to 8e-07 for $r_{\text{intra}2}$, from 5e-07 to 2e-05 for values of c , and from 1 to 35 for values of M .

From each simulation, we computed a set of summary statistics. In order to fit the diversity of simulated populations to the observed dataset, we identified SNPs from simulated alignments as previously described. To calibrate diversity patterns within populations, we computed the density of SNPs along the simulated sequences and the average Nei diversity indices for the selected SNPs. We also measured linkage disequilibrium between pairs of SNPs using the D' statistic within and among populations to study the impact of recombination rates on simulated sequences. To assess the impact of species divergence on the simulated data, we recorded the density of selected SNPs across the whole alignment, as well as the percentage of polymorphic sites in the alignment that were fixed in each population, and we then used H_{ST} to estimate differentiation between the two simulated populations.

We used standard ABC procedures to infer appropriate values for parameters of interest [69]. The Euclidean multivariate distance between the observed and simulated normalized summary statistics was calculated, and sets of parameters were accepted and stored if this distance was less than a tolerance threshold, defined so that a small fraction of simulations (less than 1%) were accepted for the estimation step. When comparing the fit of our various models to the data, we used this same rejection threshold and determined the best model using Bayes factors. Posterior distributions of parameters of interest were summarized using average values.

Intra-specific phylodynamic model

The main genotypes that we observed in *B. burgdorferi* s.s. and *B. garinii*, which were identified by both the phylogenetic and the STRUCTURE approaches, likely emerged a long time ago. In order to study how diversity is currently maintained in our sampling sites, we developed an epidemic model that took pathogens

diversity into account. More precisely, we explored an individual based model in which the infection status for a number of different bacterial genotypes (N_{bg}) was monitored over time in hosts and vectors. Our aim was to identify epidemiological settings that would allow for the maintenance of the observed level of diversity within each species.

In this model, one iteration represented one year. Two host populations of equal and constant sizes, N_{hosts} , were simulated to represent the reservoir of infection-causing bacteria at the Munster and Guebwiller sites. Each host population came into contact with a population of nymphs, with constant and equal populations sizes N_{ticks} . Due to the complete turnover in nymphs that happens each year, larvae were not explicitly included in the model but implicitly appeared at each generation as a source of new susceptible ticks that are coming in contact with hosts. Likewise, adults were not included, as they occur at lower densities and their preferred host are large mammals, which are not known to be a reservoir of the *B. burgdorferi* species complex. At each iteration of the model, each host was exposed to a variable number of nymphs, this number followed a Poisson distribution with parameter $N_{\text{ticks}}/N_{\text{hosts}}$. The bacterial genotype (or lack thereof) within each tick that fed on each host was recorded to obtain the number of contact between each individual host and each bacterial genotype. A host became infected by the i^{th} bacterial genotype according to a binomial distribution which had as parameters the number of contacts this host had with the i^{th} genotype and the probability I_{HTi} of infection occurring during each contact. Infections were persistent, so that an infected host retained this status for its whole life. Then, the entire tick population was completely replaced with susceptible individuals. These new ticks represented uninfected larvae who, in the current iteration would, feed, become infected, molt into nymphs and then infect hosts during the next iteration. Next, each tick thus fed on a host chosen at random in the population. For each of the N_{bg} different bacterial genotypes, ticks became infected according to a Bernoulli distribution which took as parameters the probability I_{HTi} of infection per contact if the host was infected, and 0 if the host was uninfected. Later, N_{wig} host individuals were exchanged between the two host populations following a Poisson distribution with parameter $F_{\text{wig}}/N_{\text{hosts}}$ where F_{wig} is a scaling parameter. Finally, a proportion R_{hosts} of each host population was renewed with susceptible individuals.

An important criterion in the parameterization of our model was that the number of genotypes present in the model at equilibrium (taking into account genetic drift and migration) was equal the number of observed genotypes in our *B. burgdorferi* s.s. and *B. garinii* samples. Therefore, for each simulation, the model was run for 500 iterations in order to allow enough time to reach this equilibrium. We began each simulation with the assumption that each host, was infected by a single bacterial genotype, and that bacterial genotypes occurred at equal frequencies within and among populations. In such an epidemiological model, one of the two main outcomes is expected: either more pathogens are lost during population replacement than are created by new infections, which leads to the extinction of pathogens; or the number of new infections per iteration is much higher than the number of individuals lost during replacement, which leads to high pathogen prevalence and the maintenance of most genotypes in co-infected individuals. Here, we focused on sets of parameters that would enable the creation of intermediate situations. In order to identify these situations, we started simulations with a high number of genotypes, i.e. $N_{bg} = 20$; assumed strong rates of transmission during contact between hosts and ticks, $I_{HTi} = I_{HTj} = 0.8$; and, as high $N_{\text{ticks}}/N_{\text{hosts}}$ ratios lead to explosive pathogen dynamics, we assumed that $N_{\text{ticks}}/N_{\text{hosts}} = 1$. To explore the remaining parameter

space, we performed simulations using combinations of the following values: N_{hosts} set at either 100, 300, 500, or 900 individuals; F_{mig} set at either 0, 0.2, 0.4, 0.6, 0.8, or 1; and R_{hosts} set at values from 0.32 to 0.42 with a 0.005 step from one to the next. Fifty simulations were performed for each combination of parameters.

To evaluate the fit of the outcome of each simulation to our observed data, we compared different summary statistics using techniques similar to an ABC approach. At the end of each simulation, a certain number of ticks were randomly selected, with the goal of mimicking our real-world sampling of chromosomal sequences; the number of ticks selected corresponded to our actual sampling effort in each site. For *B. burgdorferi* s.s., 14 pathogens were sampled from 14 infected ticks from the first population (Munster) and 9 pathogens were sampled from the second population (Guebwiller). For *B. garinii*, 13 pathogens were sampled in the Munster population and 19 in the Guebwiller population. We computed the Euclidean multivariate distance between the observed and simulated values using the following normalized summary statistics: i) genotype richness among pathogens sampled in each population (R_1 , R_2) and in the two populations combined (R_T); ii) Nei diversity indices within each population (H_1 , H_2) and differentiation between populations measured with H_{ST} . Sets of parameters were accepted and stored if the distance was lower than a tolerance threshold, defined so that less than 1% of simulations were accepted. For *B. burgdorferi* s.s., the observed statistics were $R_1 = 3$, $R_2 = 2$, $R_T = 3$, $H_1 = 0.62$, $H_2 = 0.44$, and $H_{ST} = 0.09$. For *B. garinii*, observed statistics were $R_1 = 6$, $R_2 = 6$, $R_T = 8$, $H_1 = 0.80$, $H_2 = 0.78$, and $H_{ST} = 0.05$.

Results

Genomic data

We obtained 2,351,440 sequence reads in total, which corresponded to chromosomal coverage ranging from 2.2X to 27.2X, with an average of 8.5X. Raw reads were deposited to the Sequence Read Archive (SRA) database under accession SRP031890. Whole Genome Shotgun projects have been deposited in GenBank under the accessions listed in Table S1 (BioProject PRJNA223339).

The raw sequences obtained for each genome were mapped onto reference sequences of the chromosome and plasmids cp26 and lp54. For the chromosomal sequences, we obtained at least 90% coverage of at least one of the reference sequence for 59 out of the 63 genotypes. The four strains for which we did not find 90% coverage were IPT27 and IPT70 of *B. burgdorferi* s.s. and IPT95 and IPT107 of *B. garinii*, which were subsequently excluded from further chromosomal analyses. For the plasmids cp26 and lp54, we obtained 90% reference sequence coverage for 61 strains, excluding IPT95 (which was also excluded from chromosomal analyses). Strains IPT107 and IPT136 of *B. garinii* did not reach the coverage threshold for cp26 and lp54, respectively, and were therefore excluded from further analyses of those respective plasmids.

To perform SNPs-based analyses, we defined unique sets of markers based on sequence alignments. For the entire species complex, 39757, 1569, and 3658 SNPs were identified along the chromosomal, cp26, and lp54 sequences, respectively. Within species we identified i) 5342, 480, and 403 SNPs in *B. burgdorferi* s.s.; ii) 10119, 451, and 873 in *B. garinii*; and iii) 8498, 568, and 762 in *B. afzelii* (number of sites from chromosomal, cp26, and lp54 alignments, respectively).

Inter-specific phylogenomic patterns of relatedness

To illustrate global phylogenetic relationships among the sampled and reference strains at the species level, we constructed a Neighbor-Net network based on chromosomal data (Figure 2). Within this network, the taxonomic assignments of the isolates were consistent with those obtained from PCR-RFLP data (data not shown) and published species boundaries. Within this network, all *B. burgdorferi* s.s. strains included in the network form a clade that is separate from that of strain SV1, for which a unique species name, *B. finlandensis*, has been proposed [70]. Similarly, all the *B. garinii* strains included in the network were more closely related to each other than to the *B. bavariensis* PBi reference strain. However, as previously described [35], the average amount of sequence divergence was low both between *B. burgdorferi* s.s. and *B. finlandensis*, and between *B. garinii* and *B. bavariensis*, species pairs for which we measured respective average divergences around 2.2% and 2.8%.

Contiguous 1-kb windows were obtained from alignments in order to construct region-specific phylogenies and assess the monophyly of *Borrelia* species and/or broader genetic groups using different sets of genetic data (Figure 2, Table 1). Overall, the relationships within these phylogenies were consistent with those observed in the global network, and a high percentage of the windows from each replication unit supported the monophyly of the various *Borrelia* species, although this percentage was slightly lower for the cp26 plasmid (Table 1). In the case of chromosomal data, 85% of the window-based phylogenies supported a clade that contained all the *B. burgdorferi* s.s. strains. Furthermore, 97% of trees supported a larger monophyletic group that included all *B. burgdorferi* s.s. strains as well as the *B. finlandensis* SV1 strain. Similarly, 88% of phylogenies contained a monophyletic group of *B. garinii* strains and nearly 98% of trees supported a clade that included all *B. garinii* strains and the *B. bavariensis* strain PBi. Finally, 99% of phylogenetic trees supported the monophyly of the *B. afzelii* strains.

We also wanted to investigate whether the percentage of windows that did not support the monophyly of *B. burgdorferi* s.s. (15%) and *B. garinii* (12%) could simply be explained by a lack of phylogenetic information, and not the presence of recombination. However, when we generated 1-kb windows from simulated data that did not allow for recombination, the percentages of windows that did not support the monophyly of the two bacterial species were significantly lower than those observed from the actual molecular data (Figure 2).

Genomic structure among species

H_{ST} values, i.e. measures of genetic differentiation based on SNPs, were calculated for each pair of species within the *B. burgdorferi* species complex with the aim of obtaining a quantitative measurement of the degree of genetic isolation among the sampled species (Table 2). In general, we found more differentiation between species in the chromosome and lp54 data than in the cp26 data. When we used SNPs from all replication units to examine the differentiation between *Borrelia burgdorferi* s.s. and each of the two other *B. garinii* and *B. afzelii* species, we found that over 90% of SNPs showed significant differentiation ($p < 0.05$). Likewise, more than 80% of SNP sites were significantly differentiated between sequences of *B. garinii* and *B. afzelii* (data from all three replication units).

After observing this differentiation pattern, we investigated the impact of homologous recombination on the genetic diversity of the *B. burgdorferi* species complex by calculating D' linkage disequilibrium values for all pairs of SNPs, both in i) all genomes grouped together and ii) within groups of conspecific genomes,

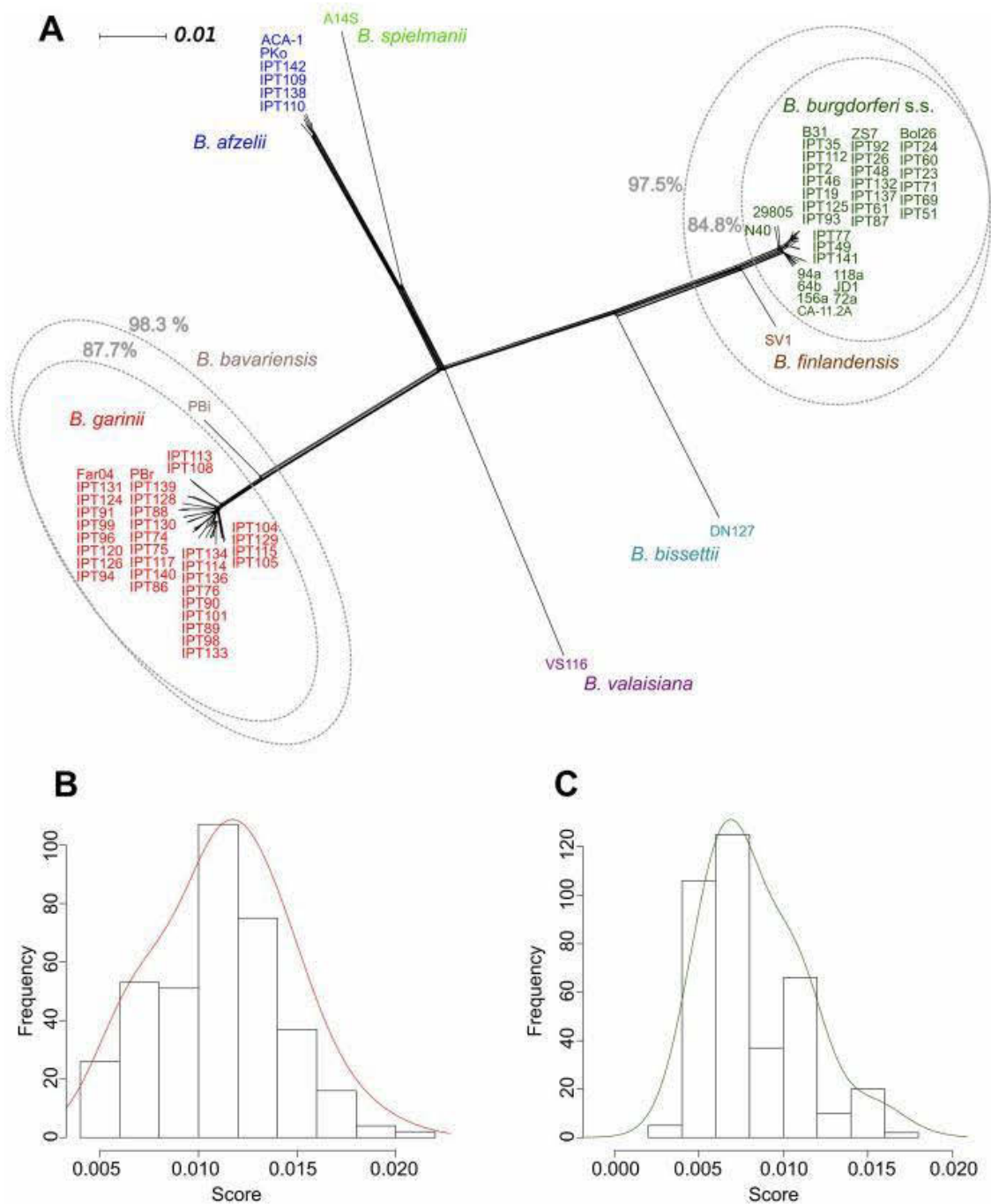


Figure 2. Phylogenetic delineation of species in the *B. burgdorferi* complex based on chromosomal sequences. (A) Neighbor-Net network based on chromosomal sequence data that illustrates phylogenetic relationships among sampled and reference strains of *Borrelia*. The network was constructed with SPLITSTREE 4 software using the Neighbor-Net method and based on a distance matrix calculated in PAUP* 4.0 b10 using a GTR+I+G model. In grey, the percentage of the 1-kb-long contiguous windows obtained from different genetic groups of *Borrelia* that support their monophyly. Distributions of expected probabilities of monophyly for strains of (B) *B. garinii* and (C) *B. burgdorferi* s.s. in the absence of recombination, based on 300 simulations of 1,000,000 genome alignments delineated into 1,000 bp windows. Continuous distributions (in red and green) were obtained using Gaussian kernels.
doi:10.1371/journal.pone.0094384.g002

Table 1. Percentage of the 1-kb-long contiguous windows obtained from different genetic groups of *Borrelia* that support their monophyly.

Genetic groups	Chromosome cp26 lp54		
<i>B. burgdorferi</i> s.s.	84.8	60.0	83.0
<i>B. burgdorferi</i> s.s. and <i>B. finlandensis</i> SV1	97.5	86.7	88.7
<i>B. garinii</i>	87.7	83.3	83.0
<i>B. garinii</i> and <i>B. bavariensis</i> PBI	98.3	90.0	96.2
<i>B. afzelii</i>	98.7	93.3	98.1
<i>B. afzelii</i> and <i>B. spielmanii</i> A145	90.2	83.3	90.6

doi:10.1371/journal.pone.0094384.t001

assigned to *B. burgdorferi* s.s., *B. garinii*, or *B. afzelii* (Table 3). As expected from our analysis of differentiation, the average D' value over the three replication units calculated for the set of all strains was high (0.92; standard error of 6.2e-06); it was higher than the values obtained for *B. burgdorferi* s.s. and *B. garinii* (0.75 and 0.81; standard errors of 7.8e-05 and 4.3e-05, respectively) but lower than that of *B. afzelii* (0.96; standard error of 2.9e-05). Nevertheless, this last value should be treated with caution due to the low sample size used to generate it. When we examined the individual replication units in the whole sample, we found a negative relationship between linkage disequilibrium and physical distance, but the decrease of D' values leveled off after 500 bp (e.g. for chromosomal data Figure 3).

Intra-specific patterns of relatedness

To more precisely illustrate the genetic relationships among the sampled and reference strains within species, we constructed Neighbor-Net phylogenetic networks based on the chromosomal data of *B. burgdorferi* s.s. and *B. garinii* (Figure 4).

Within *B. burgdorferi* s.s., most strains clustered together to form two closely related phylogenetic groups that included the published sequences of strains Bol26 and ZS7 (Figure 4). This branching pattern was unexpected, as these two published strains had been isolated in Italy and Germany, respectively, with the latter originating from a study conducted more than twenty years ago [71]. Conversely, strains IPT49, IPT77, and IPT141, which were isolated from ticks sampled from a single location in both 2003 and 2004, grouped together in a separate clade and appeared to be more closely related to North American strains of *B. burgdorferi* s.s., such as strain 94A. In fact, most of the

reticulations in this region of the network involve branches leading to American strains.

Strains assigned to *B. garinii* formed nine groups of closely related genotypes (Figure 4). These genetic groups were connected in a star-like phylogeny, with the exception of strain IPT126, which was ambiguously connected via reticulations to two different groups of strains (one including IPT75, IPT86, IPT89, IPT94, IPT117, and IPT140, and the other containing IPT76, IPT90, IPT98, IPT101, IPT114, IPT133, IPT134, and IPT136. None of the *B. garinii* strains that were sequenced for this study clustered with the reference strain *B. garinii* Far04, but we found that four of our sampled strains (IPT91, IPT96, IPT99, and IPT124) grouped with the reference strain *B. garinii* PBr, which had been isolated in Germany in the 1980's.

Finally, the network showed that the *B. afzelii* strains isolated in this study were closely related to reference strains PKo and ACA-1. However, it also showed that this group of sampled strains is not monophyletic.

Population structure within sampled species

We performed STRUCTURE analyses to describe the population structure of *B. burgdorferi* s.s. and *B. garinii* in more detail. The population structure of strains of *B. burgdorferi* s.s. suggested an optimal value of $K = 7$ populations (Figure S1). Results revealed that European strains were assigned to three main populations, which showed low levels of admixture (Figure 4). Conversely, high levels of admixture were identified among American strains. For *B. garinii*, the optimal number of populations was computed to be $K = 13$ populations, and strains from this study were assigned to 9 of those 13 populations, with strain IPT126 showing a high degree of admixture (Figure 4).

We then compared in more detail the distribution of genetic diversity within the three sampled species by calculating two estimates of the diversity statistic θ (measured per site): θ_S and θ_π . From the chromosomal data of each species, overall estimates confirmed that strains of *B. garinii* contained more genetic diversity than those of *B. burgdorferi* s.s. or *B. afzelii* (Table 4). Among our samples, in *B. burgdorferi* s.s. we found that $\theta_S = 2.7e-03$ and $\theta_\pi = 2.0e-3$. For *B. garinii*, θ_S and θ_π were 6.3e-03 and 5.5e-03, respectively and in the case of *B. afzelii*, θ_S was 4.5e-03 and θ_π was 4.6e-03.

Next, we focused on pairwise D' values for SNPs within either the *B. burgdorferi* s.s. strains or the *B. garinii* strains from both of our sampling sites. Interestingly, these values were higher and, in the case of *B. burgdorferi* s.s., considerably higher than those that we calculated from all available genomes (Table 3). For the three replication units taken together, the average D' value was 0.97 for

Table 2. Mean differentiation (H_{ST} values) of SNPs identified in different groups of *Borrelia* and the percentage of H_{ST} values that demonstrated a significant degree of differentiation ($p < 0.05$).

Groups compared	Chromosome			cp26			lp54		
	Mean	std ^a	S ^b	Mean	std ^a	S ^b	Mean	std ^a	S ^b
<i>B. burgdorferi</i> s.s. vs. <i>B. afzelii</i>	0.92	1.0e-03	95.2	0.82	7.5e-03	89.2	0.96	2.1e-03	98.0
<i>B. burgdorferi</i> s.s. vs. <i>B. garinii</i>	0.88	1.2e-03	99.7	0.83	7.5e-03	98.7	0.92	3.2e-03	99.9
<i>B. garinii</i> vs. <i>B. afzelii</i>	0.79	1.8e-03	85.0	0.73	1e-02	80.2	0.84	4.7e-03	90.3
<i>B. burgdorferi</i> s.s., Munster vs. Guebwiller	0.06	6.2e-04	0.52	0.05	2.6e-03	3.29	0.06	3.5e-03	4.6
<i>B. garinii</i> , Munster vs. Guebwiller	-0.001	4.3e-04	3.58	0.00	2.2e-03	3.85	-0.005	1.3e-03	2.6

^astandard error of the mean.^bpercentage of H_{ST} values indicating significant differentiation (%).

doi:10.1371/journal.pone.0094384.t002

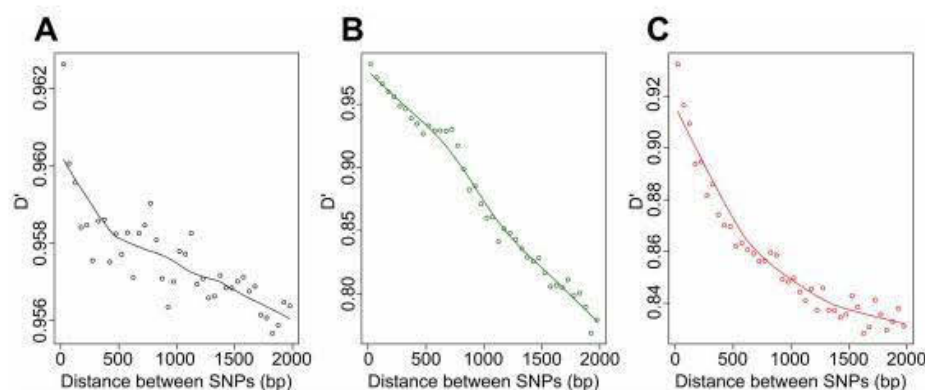


Figure 3. Relationships between D' and the physical distance between SNPs. D' values were calculated from pairs of SNPs using the chromosomal sequences of (A) all strains used in this study, (B) all strains of *B. burgdorferi* s.s., and (C) all strains of *B. garinii*. Average D' values obtained for all pairs of SNPs separated by less than 2,000 bp (using 50-bp intervals) are shown.
doi:10.1371/journal.pone.0094384.g003

B. burgdorferi s.s. and 0.81 for *B. garinii* [respective standard errors of 4.3×10^{-5} and 4.2×10^{-5}]. As for inter-specific data, we found a negative relationship between linkage disequilibrium and physical distance within species, and the decrease of D' values especially leveled off after 500 bp in *B. garinii* (Figure 3). Within both species, we found SNPs with very low D' values, indicating low levels of linkage disequilibrium with most other SNPs and a large number of these SNPs were located in the vicinity of the *ospC* gene (Figure 5).

Given that the levels of linkage disequilibrium in *B. burgdorferi* s.s. and *B. garinii* in our regional sample were higher than those calculated for the global sample, we investigated the extent of geographical isolation within each species by using SNPs frequencies to analyze the distribution of H_{ST} values between our two sampling sites (Table 2, Figure 6). At the SNP level, only a few markers (ranging from 0.52% to 4.58% of SNPs per species/replication unit combination) revealed significant isolation ($p < 0.05$) between samples obtained from Munster and those obtained from Guebwiller for either *B. burgdorferi* s.s. or *B. garinii*. However, the distribution of H_{ST} values obtained from isolates of *B. burgdorferi* s.s. was centered around a significantly higher value ($p < 2.2 \times 10^{-16}$) than the distribution of values obtained from *B. garinii*, indicating that the former species was significantly more differen-

tiated between our two study sites (Figure 6). In both species, H_{ST} values computed from SNPs located in a region of 4 kb around the *ospC* gene on cp26 deviated strongly from the rest of the distribution (Figure S2), a signal that this region is subjected to peculiar evolutionary constraints. Similarly atypical patterns could be observed in various regions of the chromosome and of the lp54 plasmid (Figure S2).

Evidence of diversifying selection along replication units

Tajima's D values were computed for 1-kb windows of sequence data obtained from each replication unit for each of the sampled species. Within each individual species, most of the mean Tajima's D values for the individual replication units were negative (Figure S3): for *B. burgdorferi* s.s., mean Tajima's D values were -1.02 , -0.13 , and -0.61 for the chromosome, cp26 and lp54 plasmid, respectively; for *B. garinii*, they were -0.58 , -0.46 , and -0.58 , respectively; and for *B. afzelii* they were -0.33 , 0.12 , -0.37 respectively.

Because strong Tajima's D values could be the result of host-driven selective pressures on the genes contained within the windows and indicate genome regions showing extensive genetic diversity, we identified the 1-kb windows within each species that had the highest absolute values. When we identified the genes that

Table 3. Mean linkage disequilibrium (D' values) at different sampling scale for pairs of SNPs within different genetic groups of *Borrelia*.

Sampling scales	Chromosome		cp26		lp54	
	Mean	std*	Mean	std*	Mean	std*
All strains	0.92	6.2e-06	0.87	2.1e-04	0.94	5.3e-05
All strains from this study	0.96	5.3e-06	0.92	1.8e-04	0.98	3.5e-05
All <i>B. burgdorferi</i> s.s. strains	0.75	7.9e-05	0.76	9.2e-04	0.66	1.2e-03
<i>B. burgdorferi</i> s.s. strains from this study	0.97	4.3e-05	0.95	5.9e-04	0.93	1.1e-03
All <i>B. garinii</i> strains	0.81	4.3e-05	0.80	9.8e-04	0.73	5.5e-04
<i>B. garinii</i> strains from this study	0.82	4.2e-05	0.81	9.9e-04	0.74	5.5e-04
All <i>B. afzelii</i> strains	0.96	2.9e-05	0.92	6.1e-04	0.95	4.6e-04
<i>B. afzelii</i> strains from this study	0.98	2.9e-05	0.97	4.7e-04	0.97	2.8e-04

doi:10.1371/journal.pone.0094384.t003

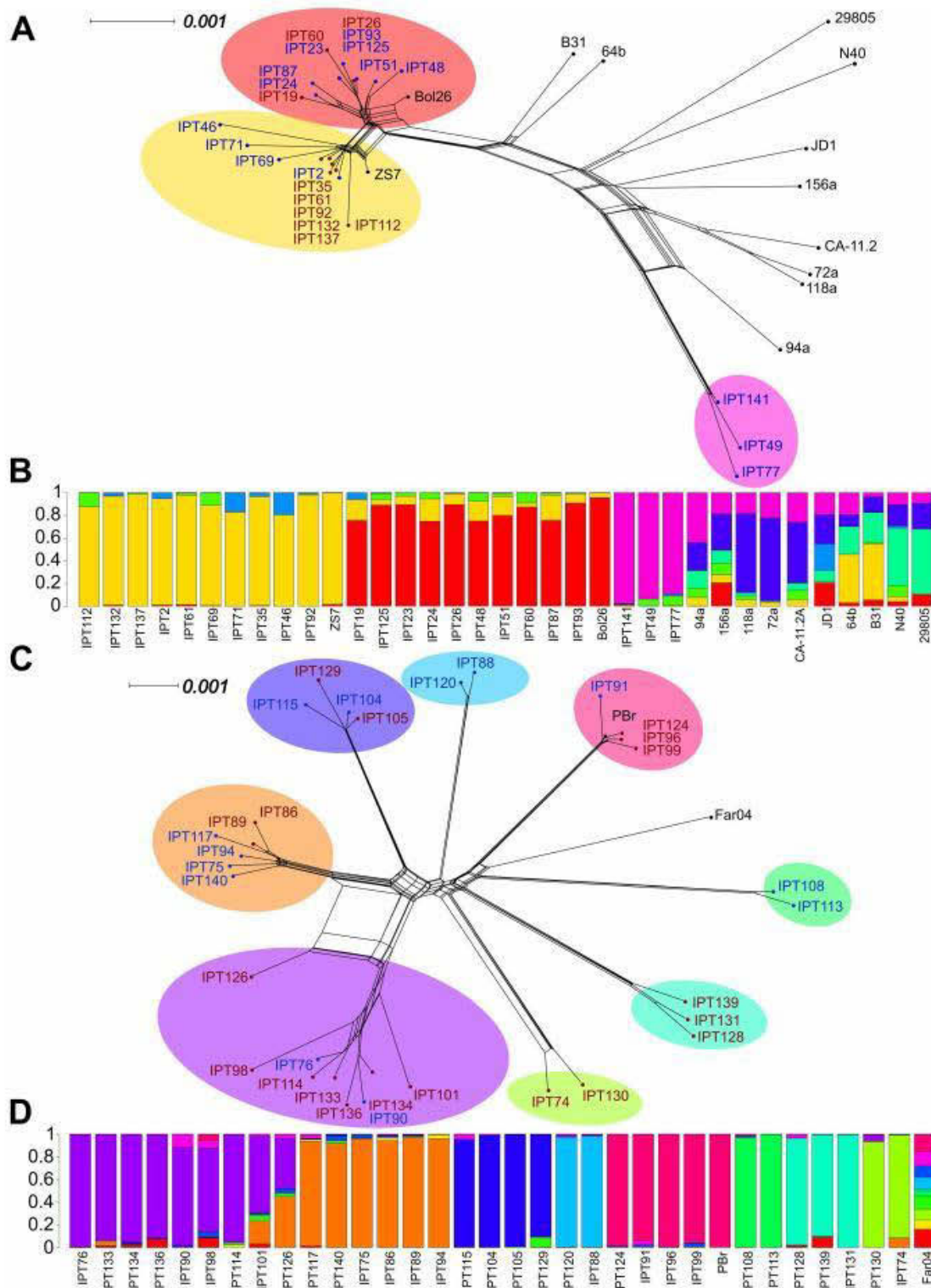


Figure 4. Phylogenies and population structures of *B. burgdorferi* s.s. and *B. garinii*. Neighbor-Net networks based on chromosomal data reveal the phylogenetic relationships among sampled and reference strains of (A) *B. burgdorferi* s.s. and (C) *B. garinii*. Networks were constructed with SplitsTree 4 software and were based on a distance matrix calculated in PAUP* 4.0 b10 using a GTR+I+G model. Strain names are surrounded in function of the main population they were assigned to using Structure; the results of the best run of STRUCTURE v2.3.4 are shown for (B) *B. burgdorferi* s.s. (K=7) and (D) *B. garinii* (K=13). Analyses were based on all polymorphic sites present in at least 90% of strains identified using multiple alignments of chromosomal data, assuming correlations among linked loci and allowing admixture among potential populations. doi:10.1371/journal.pone.0094384.g004

were present within the selected windows, we found that, compared to the total number of annotated genes for each of the three replication units (chromosome, cp26, and lp54), the genes within these highlighted windows contained a significantly lower proportion of cytoplasmic protein-encoding genes and, instead, were significantly enriched in surface lipoprotein-encoding genes (Fisher's exact test, $p < 0.05$). More precisely, genes that encoded cytoplasmic proteins (identified through SLEP analyses) only represented 55% of genes with high Tajima's D values even though they made up 67% of the genes within the total proteome encoded by the chromosome, cp26, and lp54. Genes annotated as encoding for membrane-associated proteins made up 23% of genes with high Tajima's D values and 22% of genes in the studied proteome, while genes described as encoding exported proteins represented 4% and 5% of those respective groups. Finally, 18% of the genes with high Tajima's D values were annotated as lipoprotein-encoding genes, which make up only 6% of genes in the proteome as a whole. Another observation adds weight to the unique position of lipoprotein-encoding genes in this sample is that the only gene found in high-D-value windows in all three *Borrelia* species was the *ospC* lipoprotein gene.

Phylogenetic analysis of the *ospC* gene and flanking regions

In order to more thoroughly investigate the reasons for which the region around the *ospC* gene produced peculiar results in our analyses of linkage disequilibrium, genetic differentiation, and Tajima's D values, we constructed phylogenetic networks of the *ospC* gene and flanking regions (Figure 7). The network obtained from the alignment of *ospC* sequences showed a star-like structure, with most external branches including a small number of genotypes (Figure 7C). When we examined either the 2,000 bp upstream or the 2,000 bp downstream of the *ospC* gene, the structure became more resolved (Figures 7B and 7D). Nevertheless, all of these networks differed markedly from the one based on the chromosomal alignment (Figure 2). Most notably, in these networks using cp26 data, *B. burgdorferi* s.s. did not form an obvious clade. Its constituent strains were distributed into different

subgroups that were linked by large reticulations to strains of either *B. afzelii* or *B. spielmanii*. Furthermore, *B. finlandensis* strain SV1 was included within a group of *B. burgdorferi* s.s. strains. Conversely, *B. burgdorferi* s.s. strains IPT49, IPT77, and IPT141, for which the phylogenetic analysis based on chromosomal data had revealed only a distant relationship to other *B. burgdorferi* strains, were still grouped together but were embedded among other *B. burgdorferi* s.s. strains. When we constructed networks based on sequence data further upstream or further downstream of the *ospC* gene, the patterns became more consistent with the phylogenetic signal obtained from the chromosome (Figure 7A and 7E).

Long-term coalescent based model

We developed coalescent-based models to estimate the impact of various divergence times, effective population sizes, and intra-specific and inter-specific recombination rates on the evolutionary history of *B. burgdorferi* s.s. and *B. garinii*. First, we checked that the distributions of summary statistics obtained from our accepted simulations contained the statistics generated from our observed data (data not shown). To obtain quantitative support for our inference, we used Bayes factors to compare different models that had different constraints applied to the parameters of interest (Table S4). As a result of this comparison, we chose the M3 model for further analysis. From posterior distributions of the parameters of interest (Figure S4), we computed average values for each parameter and scaled the different estimates to yield interpretable results. The resulting estimates of the θ parameter were $8.2e-04$ for *B. burgdorferi* s.s. and $1.5e-03$ for *B. garinii*, while the effective population sizes of the two species (assuming a mutation rate of $1e-07$ per site per generation) were approximately 8200 and 14500 respectively. The inter-specific recombination rates, $r_{inter} = r_{inter2}$, was estimated to be $3.1e-09$, which was 50 times lower than the intra-specific recombination rates, $r_{intra1} = r_{intra2} = r_{intra3} = 1.7e-07$. The ratio of the recombination rate to the mutation rate was approximately 1.7. The model estimated that the two species shared a common ancestor around 490,000 generations ago. The ancestral population size estimate should be treated with caution

Table 4. Chromosomal genetic diversity and Tajima's D values.

Species	θ_s		θ_a		Tajima's D	
	Mean	std*	Mean	std*	Mean	std*
All <i>B. burgdorferi</i> s.s. strains	4.2e-03	9.4e-05	3.1e-03	1.6e-04	-1.02	4.1e-02
<i>B. burgdorferi</i> s.s. strains from this study	2.7e-03	7.2e-05	2.0e-03	9.3e-05	-0.80	5.9e-02
All <i>B. garinii</i> strains	6.6e-03	8.7e-05	5.5e-03	8.7e-05	-0.58	2.1e-02
<i>B. garinii</i> strains from this study	6.3e-03	8.6e-05	5.5e-03	8.9e-05	-0.44	2.1e-02
All <i>B. afzelii</i> strains	4.0e-03	1.1e-04	3.6e-03	1.2e-04	-0.33	3.9e-02
<i>B. afzelii</i> strains from this study	4.4e-03	1.3e-04	4.5e-03	1.4e-04	0.18	7.1e-02

*standard error of the mean.

To estimate the diversity statistic θ , Watterson's θ_s and Tajima's θ_a were calculated using chromosomal sequence data for all 1-kb windows. Tajima's D values were calculated from these estimates; the mean and standard error for each group are shown.

doi:10.1371/journal.pone.0094384.t004

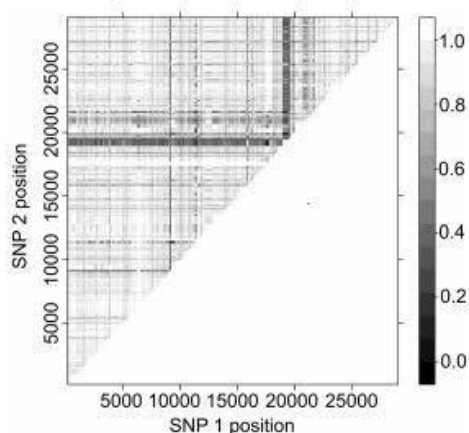


Figure 5. Linkage disequilibrium along the cp26 plasmid. The axes are the positions on the plasmid of the SNPs included in each pairwise analysis. An absence of shading corresponds to a high level of linkage disequilibrium (D' close to 1) and progressively darker shading indicates decreasing values of D' . The genome region in and around the *ospC* gene is characterized by low linkage disequilibrium.
doi:10.1371/journal.pone.0094384.g005

as the posterior distribution from which it came indicates a large degree of uncertainty regarding the outcomes of the model (Figure S4).

Intra-specific phylodynamic model

We explored the properties of an individual-based epidemiological model to better understand factors that could influence the maintenance of the diversity we observed in our two sites, Munster and Guebwiller. Independently for each species, we performed 50 simulations that assumed different values for parameters of interest: the number of hosts per population, N_{hosts} ; the replacement rate of the host population, R_{hosts} ; and the migration rate between sampling sites $F_{\text{mig}}/N_{\text{hosts}}$. Of these simulations, we selected those whose summary statistics were similar to observed values (1% of the total simulations for *B. burgdorferi* s.s. and *B. garinii*). We then plotted the values of the parameters for the selected simulations in order to identify patterns of epidemiological relevance (Figure 8). Overall, larger population sizes and higher migration rates were required to maintain the diversity that we observed in *B. garinii* than were necessary for the maintenance of observed patterns in *B. burgdorferi* s.s. According to the model, host population sizes of 100 individuals would be sufficient to function as effective reservoirs for the diversity we observed in our study.

Discussion

Distribution of diversity among species of the *Borrelia burgdorferi* complex

We studied species delineations within the *B. burgdorferi* complex using genomic data from sympatric isolates and geographically independent reference sequences. Our results show that the *B. burgdorferi* species complex is composed of different genetic groups that are clearly isolated from one another (Figure 2). These results are consistent with the robust genetic boundaries that have been observed among the members of the *B. burgdorferi* species complex following analyses of concatenated MLST data [29]. Although there were few obvious inter-specific recombination events, most

Borrelia burgdorferi Population Genomics

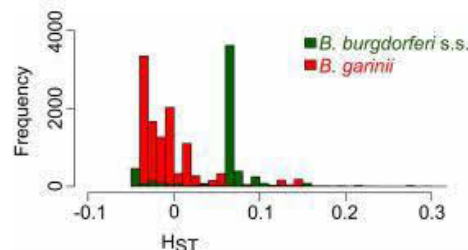


Figure 6. Genetic differentiation between sampling sites for *B. burgdorferi* s.s. and *B. garinii*. Distributions of H_{ST} values were obtained from within-species analyses of geographical genetic differentiation based on SNPs. Genetic differentiation between the two sites was significantly higher for *B. burgdorferi* s.s. than for *B. garinii*.
doi:10.1371/journal.pone.0094384.g006

phylogenies based on contiguous 1-kb windows along the three main replication units studied here (the chromosome, the cp26 plasmid, and the lp54 plasmid) showed clear delineations among the *B. burgdorferi* s.s., *B. garinii*, and *B. afzelii* strains (Table 1). This finding is further supported by the almost complete genetic differentiation that we observed in our analysis of H_{ST} values among the three species (Table 2). Finally, linkage disequilibrium measures were much higher for the species complex as a whole than within the two most frequently sampled species, *B. burgdorferi* s.s. and *B. garinii* (Table 3). Additionally, coalescent-based inference consistently revealed that within-species recombination rates for *B. burgdorferi* s.s. and *B. garinii* may be more than ~50 times higher than between-species recombination rate (Figure S4). Taken together, these data suggest that homologous recombination occurs much more frequently within than among species, which counters previous suggestions that the rate of cross-species recombination in this species complex is high [36].

Some exceptions to the monophyly of the sampled species were found: *B. burgdorferi* s.s. was non-monophyletic in 15% of the phylogenies generated here, and *B. garinii* was non-monophyletic in 12% (Figure 2). The majority of these incongruencies resulted from the inclusion of the *B. finlandensis* SV1 strain within a cluster of *B. burgdorferi* s.s. strains or the inclusion of the *B. bavariensis* PBi strain with strains of *B. garinii*. However, the average amount of sequence divergence in both cases (between *B. burgdorferi* s.s. and *B. finlandensis*, and between *B. garinii* and *B. bavariensis*) is lower than the 5% divergence usually observed among *bona fide* species [72]. This raises questions on the amount of information available to delineate these groups. Additionally, the observed percentages of windows that did not support the respective monophyly of our two study species were much higher than would be predicted by an evolutionary model lacking recombination, as our phylogenetic simulations revealed (Figure 2). Thus, recombination presents itself as a possible phenomenon that might be responsible for the conflicting phylogenetic signal among windows. However, another explanation is also likely: the inability of certain phylogenies to distinguish between and to separate these pairs of species might simply result from the incomplete sorting of closely related bacterial lineages since the time of divergence [73]. This latter hypothesis is supported by the low ratio between the inter-specific and the intra-specific rate of recombination that was observed in *B. burgdorferi* s.s. and *B. garinii*.

This sharp decrease between inter-specific and the intra-specific recombination rates in the *B. burgdorferi* species complex, shown by both the observed genomic data and the results of our coalescent-based model, could be explained by various hypotheses. From an

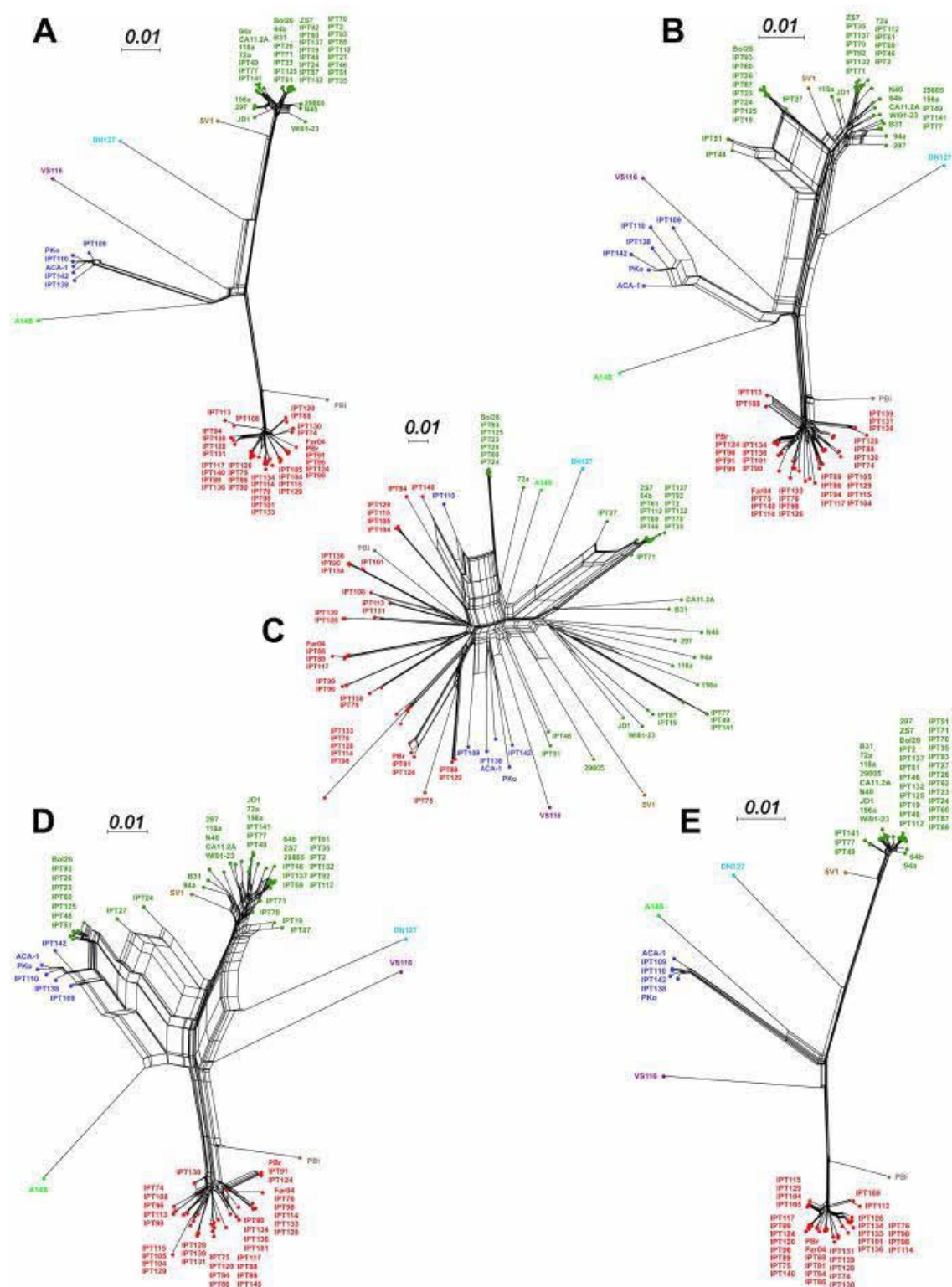


Figure 7. Phylogenetic ambiguities in the ospC gene in lineages associated with small mammals. Five Neighbor-Net networks were obtained (using p-distances and SPLITSTREE 4) from alignments of the regions including and flanking the ospC gene and are presented in the order in which they physically occur in the cp26 sequence. (A) Upstream region; (B) 2,000 bp sequence that occurs before ospC; (C) the ospC gene, (D) 2,000 bp sequence that occurs after ospC; and (E) downstream region delineated.

doi:10.1371/journal.pone.0094384.g007

ecological point of view, a lack of physical contact between genetic groups in the species complex as a result of association with different vertebrate hosts may result in divergence if mutations accumulate via genetic drift and/or selection [29], and such physical isolation may prevent the homogenization of lineages via homologous recombination. The divergence of *B. garinii* from *B. bavariensis* could be explained by this hypothesis, as it is associated with a shift in host range [18]. However, no such evidence is available to explain the separation of *B. finlandensis* from *B. burgdorferi*. We also did not find obvious evidence for homologous recombination between *B. valaisiana* and the studied *B. garinii* genomes, even though both species co-infect the same host species in our sampling location [38]. Thus, it seems that a lack of opportunity is not the only factor inhibiting a higher recombination rate. Other hypotheses could be that inter-specific recombinants are genetically less fit and do not persist in populations, or that genetic divergence among species might depress homologous recombination rates in this species complex. In this context,

looking for genes experiencing diversifying selection and presenting a high degree of divergence is particularly interesting.

Evidence of selective pressures within genomes and resulting ospC diversification

We used Tajima's D statistic to determine which genes might be affected by selective pressures, and we found that genes involved in lipoprotein production had a greater representation among genes with high Tajima's D values than would have been predicted from their prevalence in the overall proteome of each of the species studied. In this, our results concur with the genetic evidence reviewed by Brisson *et al.* [34] as well as a recent phylogenetic analysis of the ratio of non-synonymous to synonymous mutation rates [35]. Of the lipoprotein-encoding genes highlighted here, the only one identified in all three taxa was *ospC*. The high genetic diversity of the *ospC* gene is probably influenced by a combination of selective pressures. The first type of selective pressure may be a

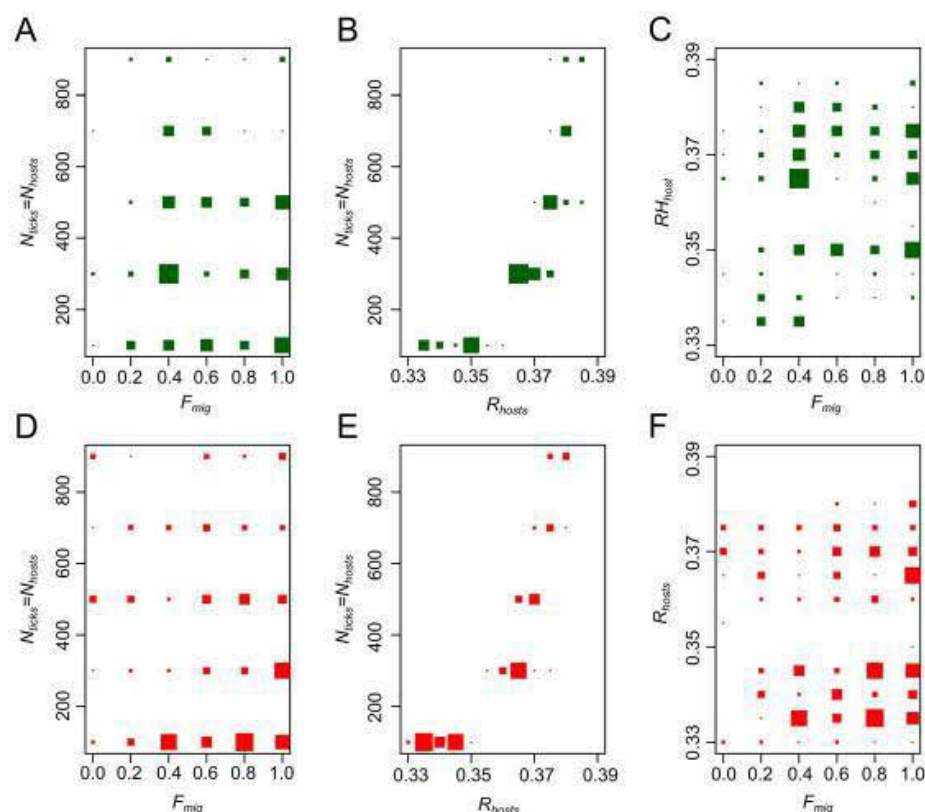


Figure 8. Distribution of the parameter values of the phylodynamic model simulations. The values of the parameters of interest for the simulations we selected (based on their similarity to our observations) are plotted. Green points correspond to simulations of *B. burgdorferi* s.s. and red points to simulations of *B. garinii* and: (A and D) F_{mig} in function of $N_{ticks} = N_{hosts}$; (B and E) R_{hosts} in function of $N_{ticks} = N_{hosts}$; (C and F) F_{mig} in function of R_{hosts} .

doi:10.1371/journal.pone.0094384.g008

consequence of the function of the lipoprotein encoded by *ospC*, which binds the tick salivary protein Salp15 [74] as well as host plasminogen [75] in order to facilitate *Borrelia* dissemination within hosts. Disruptive selective pressures on the various alleles of *ospC* could give these bacteria the ability to interact with different host species. Indeed, the multiple niche model has been suggested to be a critical driver of the diversity of these bacteria [32]. Second, the protein encoded by *ospC* has antigenic properties [76] and might thus evolve under negative frequency-dependent selective pressure [77]. Evidence of both of these types of selective pressure on *ospC* diversity was obtained from a recent study [33] that found: i) a strong association between bacterial *ospC* genotypes and the rodent hosts they infect, suggesting their limited spread in the host community; and ii) different *ospC* alleles associated with a common genetic background, suggesting the influence of diversifying selective pressure.

We did not find direct evidence of associations between *ospC* alleles and potential host species. However, our phylogenetic analysis of the flanking regions of *ospC* revealed ambiguous relationships between *B. burgdorferi* s.s. and both *B. afzelii* and *B. spielmanii*, an observation that probably indicates some degree of recombination among these taxa despite high levels of divergence. As recombination requires physical contact between genotypes, the recombinant genotypes we observed provide evidence that coinfections with *B. burgdorferi* s.s., *B. afzelii*, and *B. spielmanii* strains occurred in the past. *B. burgdorferi* s.s. can be found in both birds and small mammals, while the two other species are usually associated with small mammals. Therefore, one explanation might be that the *B. burgdorferi* s.s. lineages we sampled are infected the same hosts as *B. afzelii* and *B. spielmanii*. The reticulations in the *ospC* phylogeny suggest that inter-specific recombination may thus be involved in the emergence of new genotypes within the *B. burgdorferi* species complex (Figure 7). However, the lack of evidence of *ospC* recombination between *B. bavariensis* PBi strain and other species that infect small mammals indicates that other evolutionary scenarios are also likely. Finally, it is difficult to estimate the approximate time periods during which these inferred recombinations may have occurred. If recombinations were more recent, it would suggest that the strains of *B. burgdorferi* s.s. that we sampled frequently infect small mammals within the geographical area we studied. A future challenge for researchers will be to calibrate a molecular clock for the *ospC* region that takes into account its specific evolutionary constraints, i.e. high recombination rate and selective pressures.

When we examined the distribution of *ospC* diversity among lineages, we found closely related genomic backgrounds associated with different *ospC* alleles (Figure 7C); this was observed in both *B. burgdorferi* s.s. and *B. garinii*. We also observed a lower degree of linkage disequilibrium around the *ospC* region than in any other genome location (Figure 5). These observations could be explained by a locally higher recombination rate and/or selective pressure driven by hosts' immune systems. Considering the selective pressures likely acting on the *ospC* region, it seems probably that recombination could produce genetic combinations that would subsequently be maintained in higher numbers by diversifying selection than would persist as a result of neutral evolution or purifying selection. However, in describing multiple *ospC* alleles associated with similar genetic backgrounds and the low linkage disequilibrium between *ospC* data and other parts of the genome, our results differ substantially from those of Haven *et al.* [36] who observed a strong relationship between allelic variation in *ospC* and the rest of the genome. The authors of that study used this relationship to hypothesize a major role for negative frequency-

dependent selection in the diversification of the *B. burgdorferi* species complex, a role that our results call into question.

Genetic structure within *Borrelia burgdorferi* s.s.

In addition to the patterns observed at the *ospC* locus, our analyses also conflicted with previous observations regarding linkage disequilibrium. Within *B. burgdorferi* s.s., our random sampling at a regional scale revealed much higher linkage disequilibrium values than what can be obtained from previously sequenced genomes [36,37] (Table 3). This observation was strengthened by a STRUCTURE analysis. While the strains we sampled and sequenced, as well as the two other European genomes included here, were assigned to clearly delineated populations, genomes obtained from North American strains showed marked evidence of admixture (Figure 4B). This difference was also visible in the phylogenetic network of *B. burgdorferi* s.s. strains, in which reticulation mostly involved branches leading to American strains (Figure 4A). Originally, the sequenced American strains were "chosen to cover a large fraction of the genetic and geographic diversity" within *B. burgdorferi* s.s. and such a choice might lead the emergence of peculiar diversity patterns [37]. For example, in an analysis that used these American strains, Haven *et al.* described a positive relationship between linkage disequilibrium and physical distance [36], which was counter-intuitive [34]. However, the analysis of our local sample revealed the expected negative relationship between the two statistics (Figure 3). Additionally, the restriction of high values of linkage disequilibrium to genomic areas distant from less than a kilobase that we observed is consistent with current knowledge on the length of DNA fragment exchanged by homologous recombination [34]. The observed difference in linkage disequilibrium patterns between our regional samples and the whole sample, including Schutzer *et al.* genome sequences [37], may be compatible with the hypothesis of an epidemic population structure in *B. burgdorferi* s.s., even though this hypothesis has been previously criticized [36]. The high degree of linkage disequilibrium measured here is probably due to the prevalence in our sample of a small number of bacterial genotypes that had high transmission success, and an analysis of a collection of genomes that optimizes diversity might show more evidence of genome-wide recombination, a pattern that would be coherent with that given by Maynard Smith *et al.* in their seminal paper [13]. In this context, it would be interesting to determine whether the measured difference in linkage disequilibrium measures is due to either a difference in sample selection based on *a-priori* genomic information, or a difference in the geographic scale of sampling. Indeed it has been described in *B. afzelii* that different recombinant genotypes were isolated from independent locations [78].

Differences between the genetic structures of *B. garinii* and *B. burgdorferi* s.s. and epidemiological considerations

Linkage disequilibrium values within *Borrelia garinii* genomes sampled in Alsace were lower than those measured within strains of *B. burgdorferi* s.s. from the same locations (Table 3). Nevertheless, the STRUCTURE analysis did not reveal extensive evidence of admixture, with the exception of strain IPT126 (Figure 4C). Further, coalescent-based modeling suggested that recombination rates were similar in the two species. Assuming the same mutation rate for each species and considering only the strains we sampled for this study, we inferred a within-species ratio of recombination rate to mutation rate of 1.7 in *B. burgdorferi* s.s. and *B. garinii*, figures that were in close agreement with a previous estimation based on three pairwise comparisons of related genomes (using chromosome, cp26 and lp54 data) within *B. burgdorferi* s.s. [36]. We

observed clearer differences between *B. burgdorferi* s.s. and *B. garinii* from our analysis of population structure, in which we estimated effective population sizes and analyzed genetic differentiation between Munster and Guebwiller. As expected from our measurements of diversity, our coalescent models suggested that the effective population size of *B. burgdorferi* s.s. was lower than the effective population size of *B. garinii*. Moreover, differentiation measures, which are inversely proportional to migration rate, were higher for *B. burgdorferi* s.s. than for *B. garinii*.

However, these measures, by themselves, may not directly help in understanding the observed diversity patterns. The phylogenetic relationships between our sequences and reference sequences, which were isolated years ago in different areas, suggested that the diversification of the observed lineages of *B. burgdorferi* s.s. and *B. garinii* greatly preceded the establishment of the studied transmission networks. In order to more directly investigate current patterns of diversity, then, we implemented a simple phylodynamic model, which did not examine the diversification phase of the different genotypes but focused only on the maintenance of diversity. This model confirmed that the maintenance of the observed level of diversity requires a higher host population size and higher host migration rates in *B. garinii* than in *B. burgdorferi* s.s. These results are similar to those obtained from MLST data in a comparison of the genetic structures of rodent-associated *B. afzelii* and bird-associated *B. garinii* at a broader geographic scale [30]. The phylodynamic model thus provided this study with two important benefits. From an applied point of view, the phylodynamic information about the genetic structure of *B. burgdorferi* s.s. and *B. garinii* was coherent with the phylogenetic signal identified in the *ospC* gene region and both sets of results support the hypothesis that small mammals might be the reservoir of the *B. burgdorferi* s.s. strains studied here. From a purely academic point of view, a simple phylodynamic model that does not assume hypotheses about selective pressures contrary to multiple-niche model and negative frequency-dependent selection, allows researchers to explain the maintenance of numerous bacterial lineages with realistic host population sizes. We thus advocate that further efforts are required to incorporate explicit epidemiological constraints into evolutionary models in order to study the genome of both tick-borne and host associated strains and obtain further insight into the evolution of the *B. burgdorferi* species complex.

Supporting Information

Figure S1 Mean likelihood values of STRUCTURE runs. STRUCTURE analyses were conducted using K values ranging from 1 to 15; for each value, the analysis was repeated five times and the mean likelihood value is plotted here. Analyses were based on all the polymorphic sites identified in at least 90% of *B. burgdorferi* and *B. garinii* strains using multiple chromosomal alignments. For each species, analyses were performed assuming correlations among linked loci and allowing admixture among potential populations. Green points correspond to results for *B. burgdorferi* s.s. and red points to results for *B. garinii*. (TIFF)

Figure S2 Genetic differentiation between isolates from Munster and Guebwiller for *B. burgdorferi* s.s. and *B. garinii*. The distributions of H_{ST} values were obtained from within-species comparisons of SNPs between isolates from Munster and Guebwiller for *B. burgdorferi* s.s. (left column) and *B. garinii* (right column). The sequence alignments used in the comparisons were obtained from (A and B) the chromosome, (C and D) the cp26 plasmid, and (E and F) the lp54 plasmid. Red

circles correspond to values that showed a significant level of differentiation ($p < 0.05$).

(TIFF)

Figure S3 Distribution of Tajima's D values obtained from contiguous 1-kb windows for the three species of *Borrelia* included in this study. Values were calculated for A) chromosomal, B) cp26, and C) lp54 alignments. Green lines indicate the mean value of Tajima's D. Blue lines represent the mean of Tajima's D values calculated from chromosomal data and are replicated in each plot. Red lines indicate the 5th and 95th percentiles of chromosomal Tajima's D values for each species.

(TIFF)

Figure S4 Posterior distributions of parameters of the coalescent model M3. Distributions were obtained from the simulations, of the coalescent model M3, we selected (based on their similarity to our observations); (A) $N_1 = c_1 * N$, the population size of *B. burgdorferi* s.s., (B) $N_2 = c_2 * N$, the population size of *B. garinii*, (C) $N_3 = c_3 * N$ the ancestral population size, (D) the recombination rate within the *B. burgdorferi* s.s. species r_{intra1} , (E) the recombination rate within the *B. garinii* species r_{intra2} , (F) the recombination rate within the ancestral population r_{intra3} , (G) the inter-specific recombination rate r_{inter1} , (H) the inter-specific recombination rate r_{inter2} , (I) $M * (c_1 * N + c_2 * N)$ the time backward until the two populations merged.

(TIFF)

Table S1 Description of the 63 strains of the *B. burgdorferi* species complex that were isolated and sequenced in this study. ^A: M: Adult male; F: Adult female; N: Nymph. All strains were isolated from *Ixodes ricinus* ticks.

(DOC)

Table S2 Percentage of the length of reference sequences onto which raw sequences were mapped for each strain of the *B. burgdorferi* species complex examined in this study. Reference sequences came from i) the chromosome (Chr), the circular plasmid cp26, and the linear plasmid lp54 of *B. burgdorferi* s.s. strain B31; ii) the chromosome of *B. bavariensis* strain PBi and the cp26 and lp54 plasmids of *B. garinii* strain Far04; iii) the chromosome, cp26 plasmid, and lp54 plasmid of *B. afzelii* strain PKo.

(DOC)

Table S3 Description of the 23 sequenced strains isolated and sequenced in previous studies and used as references in this study.

(DOC)

Table S4 Bayes factors for each pair of coalescent models. Each cell gives the ratio of the number of simulations selected for the column model to the number selected for the row model (i.e., column: row).

(DOC)

Acknowledgments

We thank Isabelle Lebert for preparing the map of the sampling sites of the study. We thank the Genoscreen team, especially Stéphanie Ferreira and Christine Hubans, for handling the sequencing process. We also thank Jessica Pearce and Lindsay Higgins for proofreading and editing our manuscript.

Author Contributions

Conceived and designed the experiments: EF XB. Performed the experiments: EF VC M. Garnier NS XB. Analyzed the data: MJ XB.

References

- Woolhouse MEJ, Gowtage-Sequeria S (2005) Host range and emerging and reemerging pathogens. *Emerg Infect Dis* 11: 1842–1847.
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* 4: 981–994.
- Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303.
- Hanage WP, Fraser C, Spratt BG (2006) The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol* 239: 210–219.
- Hill W, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8: 269–294.
- Gilligan C, van den Bosch F (2008) Epidemiological models for invasion and persistence of pathogens. *Annu Rev Phytopathol* 46: 385–418.
- Wright S (1931) Evolution in Mendelian populations. *Genetics* 16: 97–159.
- Moran P (1958) Random processes in genetics. *Math Proc Cambridge Philos Soc* 54: 60–.
- Volz EM, Koelle K, Bedford T (2013) Viral phylodynamics. *PLoS Comput Biol* 9: e1002947.
- Taberlet P, Fumagalli L, Wust-Sauney A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Mol Ecol* 7: 453–464.
- Haydon DT, Cleaveland S, Taylor LH, Laurensen MK (2002) Identifying reservoirs of infection: a conceptual and practical challenge. *Emerg Infect Dis* 8: 1468–1473.
- Koelle K, Rasmussen DA (2012) Rates of coalescence for common epidemiological models at equilibrium. *J R Soc Interface* 9: 997–1007.
- Maynard Smith J, Smith N, O'Rourke M, Spratt BG (1993) How clonal are bacteria? *Proc Natl Acad Sci U S A* 90: 4384–4388.
- Hudson RR, Kaplan NL (1986) On the divergence of alleles in nested subsamples from finite populations. *Genetics* 113: 1057–1076.
- Balloux F (2010) Demographic influences on bacterial population structure. *Bact Popul Genet Infect Dis*: 103–120.
- Rizzoli A, Hauffe H, Carpi G, Vourch G, Neteler M, et al. (2011) Lyme borreliosis in Europe. *Euro Surveill* 16: 1–8.
- Kurtenbach K, Peacey M, Rijpkema SG, Hoodless AN, Nuttall PA, et al. (1998) Differential transmission of the genospecies of *Borrelia burgdorferi* sensu lato by game birds and small rodents in England. *Appl Environ Microbiol* 64: 1169–1174.
- Margos G, Vollmer SA, Cornet M, Garnier M, Fingerle V, et al. (2009) A new *Borrelia* species defined by multilocus sequence analysis of housekeeping genes. *Appl Environ Microbiol* 75: 5410–5416.
- Duhška L, Literák I, Kocianova E, Taragelova V, Sychra O (2009) Differential role of passerine birds in distribution of *Borrelia* spirochetes, based on data from ticks collected from birds during the postbreeding migration period in Central Europe. *Appl Environ Microbiol* 75: 596–602.
- Hanincová K, Taragelová V, Koci J, Schäfer SM, Hails R, et al. (2003) Association of *Borrelia garinii* and *B. valaisiana* with songbirds in Slovakia. *Appl Environ Microbiol* 69: 2825–2830.
- Taragelová V, Koci J, Hanincová K, Kurtenbach K, Derdákova M, et al. (2008) Blackbirds and song thrushes constitute a key reservoir of *Borrelia garinii*, the causative agent of borreliosis in Central Europe. *Appl Environ Microbiol* 74: 1289–1293.
- Humair P, Postic D, Wallich R, Gern L (1998) An avian reservoir (*Turdus merula*) of the Lyme borreliosis spirochetes. *Zentralblatt für Bakteriologie* 186: 521–538.
- Humair PF, Rais O, Gern L (1999) Transmission of *Borrelia afzelii* from *Apodemus* mice and *Citellus* voles to *Ixodes ricinus* ticks: differential transmission pattern and overwintering maintenance. *Parasitology* 118: 33–42.
- Hu CM, Humair P-F, Wallich R, Gern L (1997) *Apodemus* sp. rodents, reservoir hosts for *Borrelia afzelii* in an endemic area in Switzerland. *Zentralblatt für Bakteriologie* 185: 558–564.
- Hanincová K, Schäfer SM, Etti S, Sewell H-S, Taragelová V, et al. (2003) Association of *Borrelia afzelii* with rodents in Europe. *Parasitology* 126: 11–20.
- Brisson D, Dykhuizen DE (2006) A modest model explains the distribution and abundance of *Borrelia burgdorferi* strains. *Am J Trop Med Hyg* 74: 615–622.
- Hanincová K, Kurtenbach K, Diuk-Wasser M, Brei B, Fish D (2006) Epidemic spread of Lyme borreliosis, northeastern United States. *Emerg Infect Dis* 12: 604–611.
- Richter D, Spielman A, Komar N, Matuschka F-R (2000) Competence of American robins as reservoir hosts for Lyme disease spirochetes. *Emerg Infect Dis* 6: 133–138.
- Margos G, Vollmer SA, Ogden NH, Fish D (2011) Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. *Infect Genet Evol* 11: 1545–1563.
- Vollmer SA, Bormane A, Dinnis RE, Seelig F, Dobson ADM, et al. (2011) Host migration impacts on the phylogeography of Lyme borreliosis spirochaete species in Europe. *Environ Microbiol* 13: 184–192.
- Brisson D, Dykhuizen DE, Ostfeld RS (2008) Conspicuous impacts of inconspicuous hosts on the Lyme disease epidemic. *Proc R Soc Biol Sci* 275: 227–235.
- Brisson D, Dykhuizen DE (2004) ospC diversity in *Borrelia burgdorferi*: different hosts are different niches. *Genetics* 168: 713–722.
- Jacquot M, Bisseux M, Abrial D, Marsot M, Ferquel E, et al. (2014) High-throughput sequence typing reveals genetic differentiation and host specialization among populations of the *Borrelia burgdorferi* species complex. *PLoS One* 9: e88381.
- Brisson D, Drecktrah D, Eggers CH, Samuels DS (2012) Genetics of *Borrelia burgdorferi*. *Annu Rev Genet* 46: 515–536.
- Mongodin EF, Casjens SR, Bruno JF, Xu Y, Drabek E, et al. (2013) Inter- and intra-specific pan-genomes of *Borrelia burgdorferi* sensu lato: genome stability and adaptive radiation. *BMC Genomics* 14: 693.
- Haven J, Vargas LG, Mongodin EF, Xue V, Hernandez Y, et al. (2011) Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in *Borrelia burgdorferi*, the Lyme disease bacterium. *Genetics* 189: 951–966.
- Schutzer SE, Fraser-Liggett CM, Casjens SR, Qiu W-G, Dunn JJ, et al. (2011) Whole-genome sequences of thirteen isolates of *Borrelia burgdorferi*. *J Bacteriol* 193: 1018–1020.
- Ferquel E, Garnier M, Marie J, Bernède-Baudouin C, Baranton G, et al. (2006) Prevalence of *Borrelia burgdorferi* sensu lato and *Anaplasmataceae* members in *Ixodes ricinus* ticks in Alsace, a focus of Lyme borreliosis endemicity in France. *Appl Environ Microbiol*.
- Postic D, Assous MV, Grimont PAD, Baranton G (1994) Diversity of *Borrelia burgdorferi* sensu lato evidenced by Restriction Fragment Length Polymorphism of rrf (5S)-rrl (23S) intergenic spacer amplicons. *Int J Syst Bacteriol* 44: 743–752.
- Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5: e11147.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.
- Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.
- Wright S (1965) The interpretation of population structure by F-Statistics with special regard to systems of mating. *Evolution (N Y)* 19: 395–420.
- Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Mol Biol Evol* 9: 138–151.
- Nei M (1973) Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* 70: 3321–3323.
- Beaumont M (2005) Adaptation and speciation: what can Fst tell us? *Trends Ecol Evol* 20: 435–440.
- Team R (2011) R: A language and environment for statistical computing: 900051.
- Lewontin R (1964) The interaction of selection and linkage. *Genetics* 50: 757–782.
- Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331–341.
- Guindon S, Gascuel O (2003) A Simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696–704.
- Akaike H (1974) A new look at the statistical model identification. *Autom Control IEEE Trans* 19: 1974.
- Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20: 289–290.
- Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23: 254–267.
- Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* 21: 255–265.
- Swofford D (2003) PAUP*: phylogenetic analysis using parsimony, version 4.0 b10: 2003.
- Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci* 17.
- Gu X, Fu Y, Li W (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol*.
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406–425.
- Duthel J, Gaillard S, Bazin E, Glémin S, Ranwez V, et al. (2006) Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7: 188.
- Waterson G (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*: 256–276.

61. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437–460.
62. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
63. Giombini E, Orsini M, Carrabino D, Tramontano A (2010) An automatic method for identifying surface proteins in bacteria: SLEP. *BMC Bioinformatics* 11: 39.
64. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population Genetics. *Methods* 2035: 2025–2035.
65. Kingman J (1982) The coalescent. *Stoch Process their Appl* 13: 235–248.
66. Wiuf C, Hein J (2000) The coalescent with gene conversion. *Genetics* 155: 451–462.
67. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surv Evol Biol* 7: 44.
68. Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro H, editor. *Mammalian Protein metabolism, III*. New-York. pp. 21–132.
69. Berjorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* 19: 2609–2625.
70. Casjens SR, Fraser-Liggett CM, Mongodin EF, Qiu W-G, Dunn JJ, et al. (2011) Whole genome sequence of an unusual *Borrelia burgdorferi* sensu lato isolate. *J Bacteriol* 193: 1409–1490.
71. Schaible U, Kramer M, Museteanu C, Zimmer G, Mossmann H, et al. (1989) The severe combined immunodeficiency (scid) mouse. A laboratory model for the analysis of Lyme arthritis and carditis. *J Exp Med* 170: 1427–1432.
72. Konstantinidis KT, Ramete A, Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361: 1929–1940.
73. Castillo-Ramírez S, González V (2008) Factors affecting the concordance between orthologous gene trees and species tree in bacteria. *BMC Evol Biol* 8: 300.
74. Ramamoorthi N, Narasimhan S, Pal U, Bao F, Yang XF, et al. (2005) The Lyme disease agent exploits a tick protein to infect the mammalian host. *Nature* 436: 573–577.
75. Önder Ö, Humphrey PT, McOmber B, Korobova F, Francella N, et al. (2012) OspC is potent plasminogen receptor on surface of *Borrelia burgdorferi*. *J Biol Chem* 287: 16860–16868.
76. Wilske B, Preac-Mursic V, Jauris S, Hofmann A, Pradel I, et al. (1993) Immunological and molecular polymorphisms of *ospC*, an immunodominant major outer surface protein of *Borrelia burgdorferi*. *Infect Immun* 61: 2182–2191.
77. Qiu W-G, Bosler EM, Campbell JR, Uginé GD, Wang I-N, et al. (1997) A population genetic study of *Borrelia burgdorferi* sensu stricto from eastern Long Island, New York, suggested frequency-dependent selection, gene flow and host adaptation. *Hereditas* 127: 203–216.
78. Hellgren O, Andersson M, Råberg L (2011) The genetic structure of *Borrelia afzelii* varies with geographic but not ecological sampling scale. *J Evol Biol* 24: 159–167.

Données supplémentaires de l'article 1

Figure S1

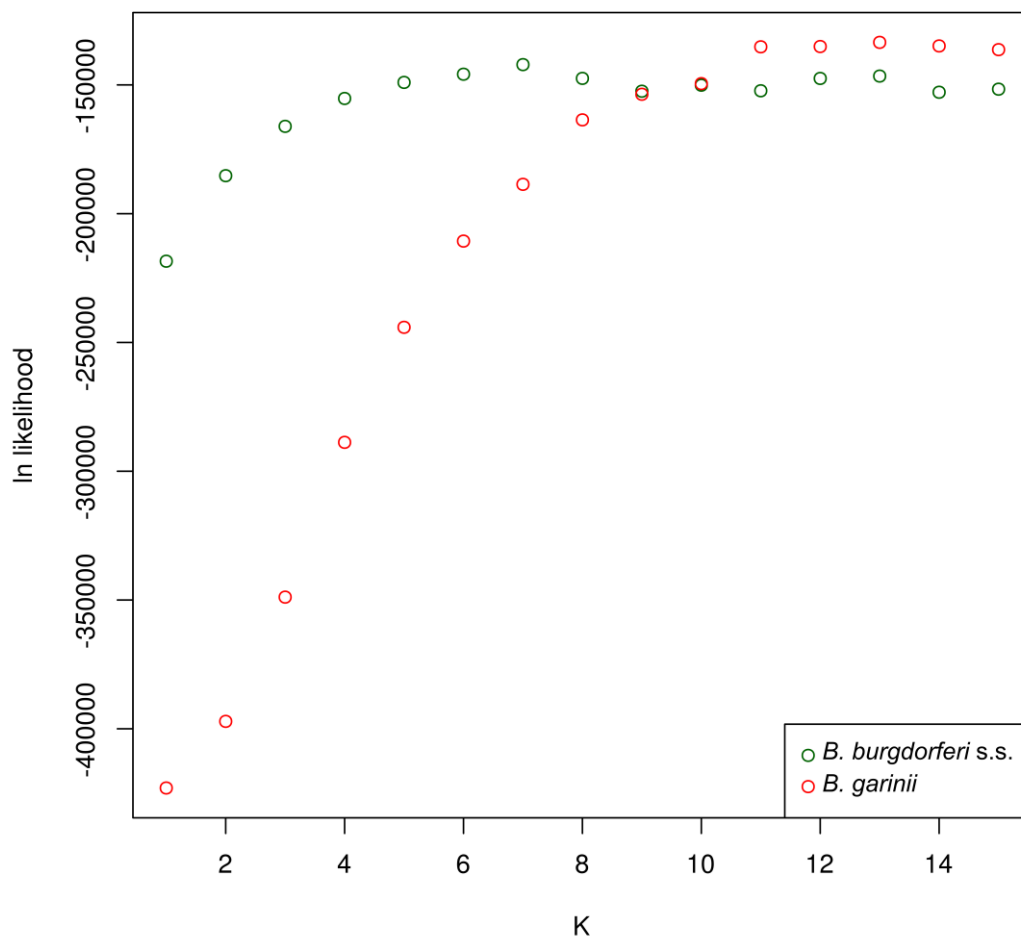


Figure S2

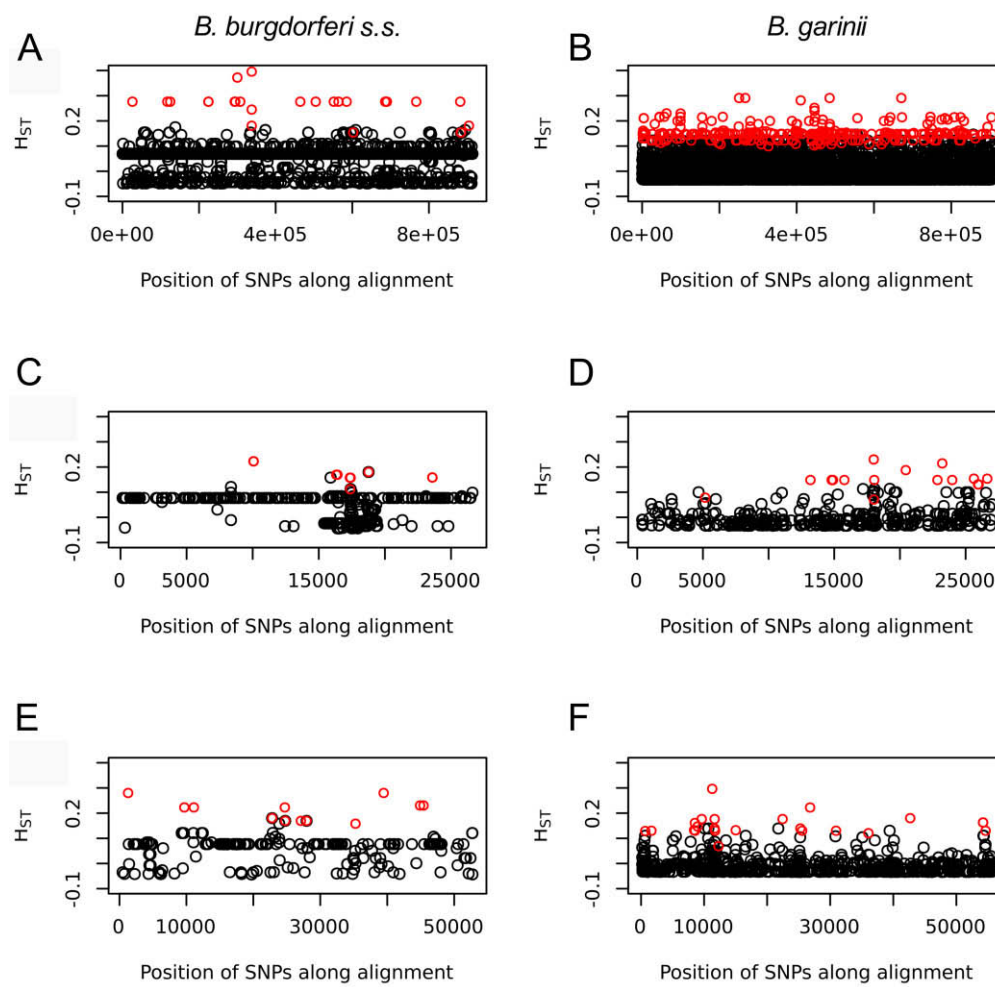


Figure S3

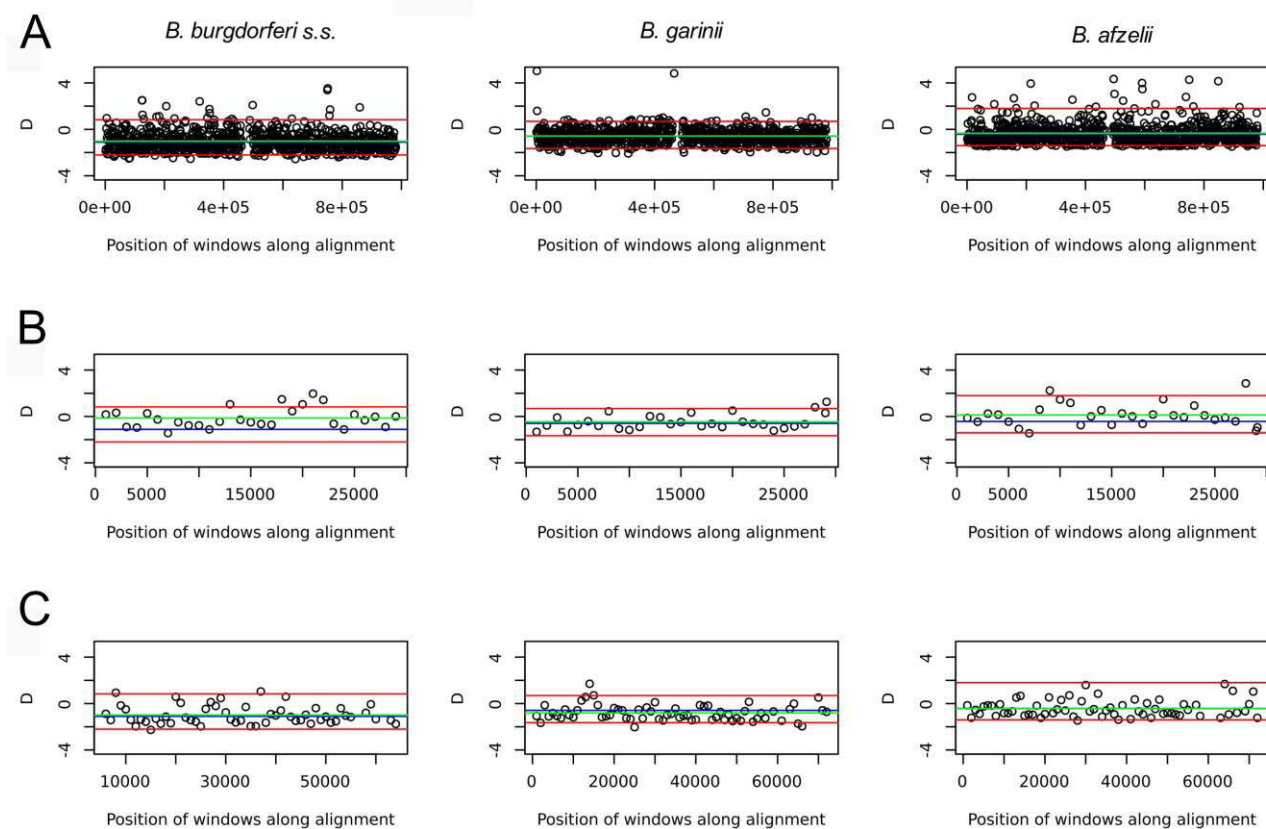


Figure S4

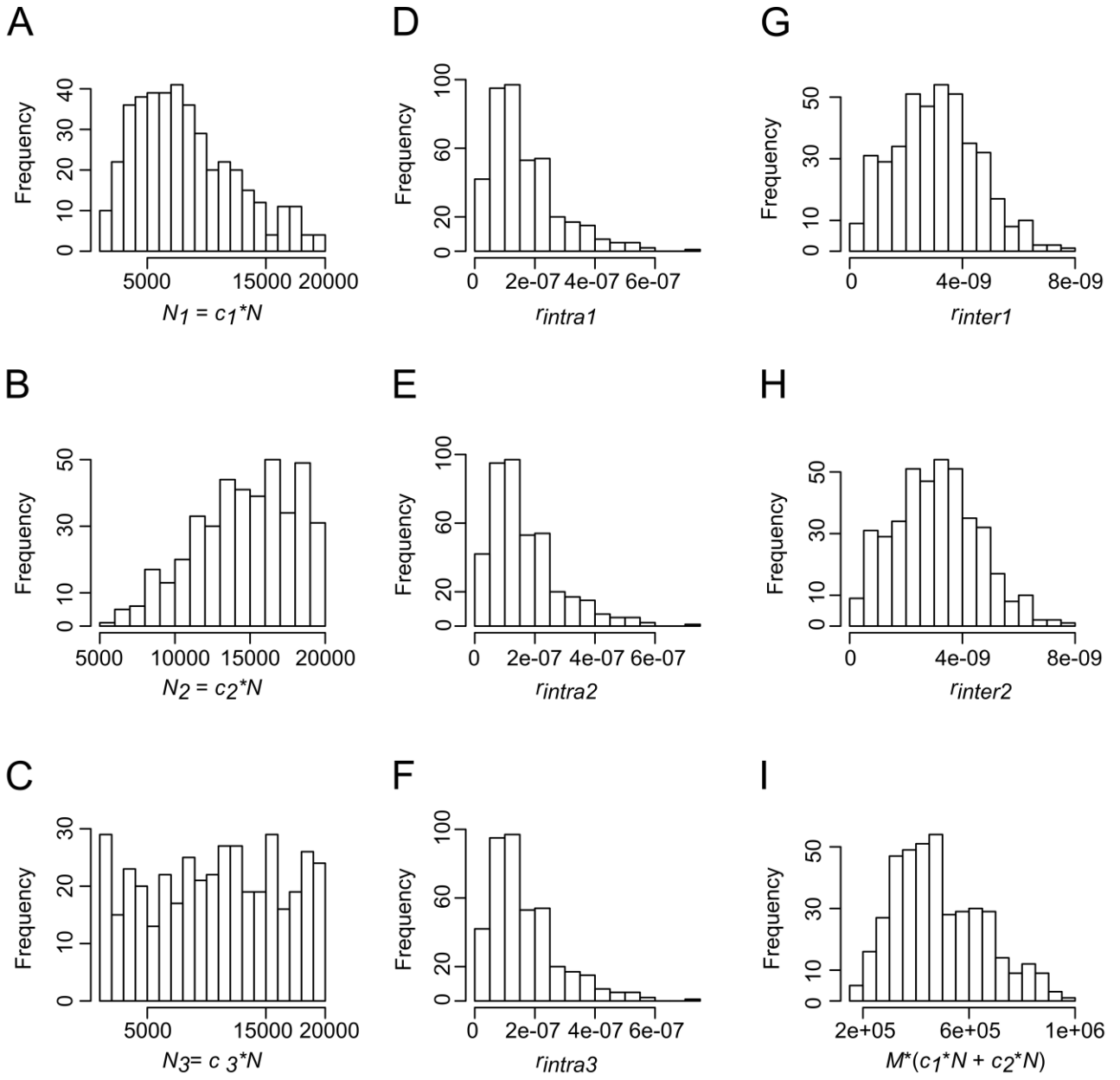


Table S1

Strain	Species	Sampling location	Development stage/sex ^A	Year
IPT74	<i>B. garinii</i>	Guebwiller	F	2004
IPT75	<i>B. garinii</i>	Munster	M	2004
IPT76	<i>B. garinii</i>	Munster	M	2004
IPT86	<i>B. garinii</i>	Guebwiller	M	2004
IPT88	<i>B. garinii</i>	Munster	M	2004
IPT89	<i>B. garinii</i>	Guebwiller	F	2004
IPT90	<i>B. garinii</i>	Munster	M	2004
IPT91	<i>B. garinii</i>	Munster	M	2004
IPT94	<i>B. garinii</i>	Munster	M	2004
IPT95	<i>B. garinii</i>	Munster	F	2004
IPT96	<i>B. garinii</i>	Guebwiller	F	2004
IPT98	<i>B. garinii</i>	Guebwiller	F	2004
IPT99	<i>B. garinii</i>	Guebwiller	F	2004
IPT101	<i>B. garinii</i>	Guebwiller	F	2004
IPT104	<i>B. garinii</i>	Munster	F	2004
IPT105	<i>B. garinii</i>	Guebwiller	F	2004
IPT107	<i>B. garinii</i>	Munster	M	2004
IPT108	<i>B. garinii</i>	Munster	M	2004
IPT113	<i>B. garinii</i>	Munster	F	2004
IPT114	<i>B. garinii</i>	Guebwiller	M	2004
IPT115	<i>B. garinii</i>	Munster	F	2004
IPT117	<i>B. garinii</i>	Munster	F	2004
IPT120	<i>B. garinii</i>	Munster	M	2004
IPT124	<i>B. garinii</i>	Guebwiller	M	2004
IPT126	<i>B. garinii</i>	Guebwiller	M	2004
IPT128	<i>B. garinii</i>	Guebwiller	F	2004
IPT129	<i>B. garinii</i>	Guebwiller	F	2004
IPT130	<i>B. garinii</i>	Guebwiller	M	2004
IPT131	<i>B. garinii</i>	Guebwiller	M	2004
IPT133	<i>B. garinii</i>	Guebwiller	F	2004
IPT134	<i>B. garinii</i>	Guebwiller	F	2004
IPT136	<i>B. garinii</i>	Guebwiller	M	2004
IPT139	<i>B. garinii</i>	Guebwiller	F	2004

Chapitre 1

Strain	Species	Sampling location	Development stage/sex ^A	Year
IPT140	<i>B. garinii</i>	Munster	M	2004
IPT2	<i>B. burgdorferi</i> s.s.	Munster	M	2003
IPT19	<i>B. burgdorferi</i> s.s.	Guebwiller	M	2003
IPT23	<i>B. burgdorferi</i> s.s.	Munster	F	2003
IPT24	<i>B. burgdorferi</i> s.s.	Munster	M	2003
IPT26	<i>B. burgdorferi</i> s.s.	Guebwiller	M	2003
IPT27	<i>B. burgdorferi</i> s.s.	Guebwiller	M	2003
IPT35	<i>B. burgdorferi</i> s.s.	Guebwiller	M	2003
IPT46	<i>B. burgdorferi</i> s.s.	Munster	F	2003
IPT48	<i>B. burgdorferi</i> s.s.	Munster	M	2003
IPT49	<i>B. burgdorferi</i> s.s.	Munster	M	2003
IPT51	<i>B. burgdorferi</i> s.s.	Munster	F	2003
IPT60	<i>B. burgdorferi</i> s.s.	Guebwiller	F	2003
IPT61	<i>B. burgdorferi</i> s.s.	Guebwiller	F	2003
IPT69	<i>B. burgdorferi</i> s.s.	Munster	N	2003
IPT70	<i>B. burgdorferi</i> s.s.	Guebwiller	N	2003
IPT71	<i>B. burgdorferi</i> s.s.	Munster	N	2003
IPT77	<i>B. burgdorferi</i> s.s.	Munster	M	2004
IPT87	<i>B. burgdorferi</i> s.s.	Munster	F	2004
IPT92	<i>B. burgdorferi</i> s.s.	Guebwiller	F	2004
IPT93	<i>B. burgdorferi</i> s.s.	Munster	M	2004
IPT112	<i>B. burgdorferi</i> s.s.	Guebwiller	M	2004
IPT125	<i>B. burgdorferi</i> s.s.	Munster	M	2004
IPT132	<i>B. burgdorferi</i> s.s.	Guebwiller	M	2004
IPT137	<i>B. burgdorferi</i> s.s.	Guebwiller	F	2004
IPT141	<i>B. burgdorferi</i> s.s.	Munster	F	2004
IPT109	<i>B. afzelii</i>	Munster	F	2004
IPT110	<i>B. afzelii</i>	Guebwiller	F	2004
IPT138	<i>B. afzelii</i>	Munster	F	2004
IPT142	<i>B. afzelii</i>	Guebwiller	M	2004

Table S2

Strain	<i>B. burgdorferi</i> s.s.			<i>B. bavariensis/B. garinii</i>			<i>B. afzelii</i>		
	B31			PBi	Far04	Far04	PKo		
	Chr	cp26	lp54	Chr	cp26	lp54	Chr	cp26	lp54
IPT74	97.64	97.73	78.37	99.06	99.88	98.34	98.57	98.09	84.04
IPT75	98.16	97.77	76.9	99.3	99.7	96.45	98.96	97.24	81.54
IPT76	97.89	97.12	77.68	99.22	99.64	97.65	98.78	97.96	84.35
IPT86	96.03	96.92	73	97.21	99.1	94.15	96.81	97.02	80.52
IPT88	97.81	99.12	78.23	99.09	100	96.95	98.66	98.48	84.02
IPT89	98.28	98.28	80.27	99.27	100	96.8	98.93	98.38	85.55
IPT90	96.93	90.38	74.92	98.53	94.52	96.72	97.93	90.12	82.34
IPT91	96.35	96.84	76.12	97.97	99.99	97.12	97.21	98.11	81.61
IPT94	97.99	94.72	80.49	99.13	96.57	96.73	98.82	95.42	84.93
IPT95	65.63	76.85	54.12	67.04	79.25	75.44	66.29	78.07	59.05
IPT96	98.17	98.99	80.68	99.38	100	97.03	99.03	98.04	85.49
IPT98	94.81	94.21	72.66	96.91	99.1	95.11	95.87	95.22	79.22
IPT99	97.37	97.81	77.94	98.82	99.76	97.52	98.2	97.42	82.5
IPT101	94.07	91.54	73.17	95.61	94.69	94.99	94.85	92.87	81.47
IPT104	98.17	99.09	79.52	99.24	99.96	97.46	98.89	98.6	84.94
IPT105	98.23	98.98	80.77	99.35	99.94	99.1	98.89	98.86	85.65
IPT107	86.18	81.63	68.58	88.2	85.47	91.37	87.03	83.18	75.44
IPT108	98.29	98.88	80.99	99.36	100	97.6	99.06	97.61	84.6
IPT113	96.98	91.34	72.36	98.44	95.81	95.65	97.81	92.78	78.69
IPT114	96.78	90.95	74.31	98.36	94.02	96.43	97.77	90.43	83.69
IPT115	96.47	96.57	77.04	97.99	98.4	97.66	97.14	94.8	84.31
IPT117	93.6	95.21	70.05	95.11	98.11	90.8	94.47	96.13	76.59
IPT120	98.03	98.67	78.59	99.29	99.83	95.9	98.85	98.35	84.89
IPT124	97.92	97.39	79.15	99.15	100	97.38	98.77	98.8	82.75

Chapitre 1

Strain	<i>B. burgdorferi</i> s.s.			<i>B. bavariensis/B. garinii</i>			<i>B. afzelii</i>		
	B31			PBi	Far04	Far04	PKo		
	Chr	cp26	lp54	Chr	cp26	lp54	Chr	cp26	lp54
IPT126	95.65	96.06	75.78	97.08	98.63	97.52	96.44	95.86	83.21
IPT128	96.52	98.67	69.11	98.1	99.71	94.6	97.47	98.16	75.53
IPT129	96.06	95.93	80.5	97.77	97.96	98.64	96.98	95.79	84.58
IPT130	97.6	95.86	75.81	99	97.04	95.68	98.52	95	83.51
IPT131	97.47	98.08	77.65	98.86	99.34	98.03	98.43	96.83	83.79
IPT133	97.35	94.41	79.52	98.67	97.01	97.93	98.23	93.96	87.56
IPT134	98.16	99	79.47	99.34	99.96	97.81	98.92	98.98	87.67
IPT136	96.36	97.3	75.06	97.96	99.47	94.96	97.19	96.77	81.59
IPT139	97.52	98.96	76.8	98.98	99.85	98.68	98.48	98.15	83.49
IPT140	97.02	98.79	78.41	98.22	99.83	96.44	97.65	98.33	84.98
IPT2	99.3	99.92	99.21	98.58	96.64	73.94	98.56	96.45	77.59
IPT19	98.98	99.22	99.73	98.2	95.39	75	98.16	96.15	76.68
IPT23	99.17	99.42	99.49	98.53	96.75	75.46	98.51	97.84	76.2
IPT24	99.27	99.31	99.42	98.46	95.9	76.19	98.45	96.7	77.34
IPT26	99.36	99.6	99.72	98.87	96.9	76.72	98.8	97.95	76.8
IPT27	84.59	94.71	94.73	83.56	90.56	69.21	83.39	91.81	71.2
IPT35	99.23	99.94	99.41	98.62	96.4	74.72	98.59	96.48	76.45
IPT46	97.23	99.51	98.62	95.71	93.4	68.82	95.68	94.3	71.52
IPT48	98.84	98.75	99.01	98.1	96.16	74.89	98	96.41	76.63
IPT49	98.15	98.14	96.53	97.33	94.11	70.99	97.22	93.36	73.68
IPT51	99.33	98.86	99.79	98.83	96.15	76	98.75	97.03	78.7
IPT60	98.23	99.54	99.33	97.21	96.73	74.62	97.28	97.85	75.58
IPT61	99.3	99.95	99.29	98.72	96.83	74.15	98.71	96.43	77.84
IPT69	96.9	99.94	97.81	96.16	96.67	71.92	96.12	96.6	75.25
IPT70	80.75	95.01	91.35	79.57	90.11	66.78	79.49	89.82	66.8

Strain	<i>B. burgdorferi</i> s.s.			<i>B. bavariensis/B. garinii</i>			<i>B. afzelii</i>		
	B31			PBi	Far04	Far04	PKo		
	Chr	cp26	lp54	Chr	cp26	lp54	Chr	cp26	lp54
IPT71	93.22	99.66	98.94	92.3	96.01	71.53	92.28	95.6	73.29
IPT77	97.62	98.58	96.88	96.48	93.33	68.64	96.59	95.11	74.24
IPT87	99.3	99.33	99.35	98.81	95.57	75.62	98.74	96.88	79.04
IPT92	99.26	99.94	99.56	98.74	96.5	77.51	98.7	96.48	78.53
IPT93	99.38	99.68	99.85	98.9	96.88	77.41	98.78	97.79	78.37
IPT112	96.65	99.33	99.29	95.66	96.26	73.91	95.65	95.83	74.71
IPT125	98.45	99.51	99.83	97.7	96.67	75.81	97.67	97.8	77.24
IPT132	99.19	99.94	99.84	98.58	96.97	75.32	98.57	96.93	78.02
IPT137	99.26	99.95	99.77	98.7	96.87	76.12	98.59	96.24	77.33
IPT141	99.19	99.49	98.41	98.91	96.55	77.93	98.9	97.59	79.24
IPT109	96.76	96.9	81.38	97.32	96.88	86.17	97.95	99.96	97.28
IPT110	93.81	95.4	78.79	94.77	94.58	84.25	95.91	99.4	97.95
IPT138	98.28	95.57	81.64	98.86	95.76	86.72	99.28	99.89	97.48
IPT142	96.3	92.17	80.31	96.9	92.24	83.5	97.88	97.98	96.5

Chapitre 1

Table S3

Strain	Genospecies	Geographical origin	Biological source	Accession Number Chr	Accession Number cp26	Accession Number lp54	Reference(s)
Far04	<i>B. garinii</i>	Denmark	Bird	ABPZ02000001-33	CP001319	CP001318	[1]
PBr	<i>B. garinii</i>	Denmark	Human	ABJV02000001-5	CP001305	CP001308	[1]
94a	<i>B. burgdorferi</i> s.s.	USA (New York)	Human	ABGK02000001-9	CP001493	CP001500	[2]
B31	<i>B. burgdorferi</i> s.s.	USA (New York)	<i>I. scapularis</i>	AE000783	AE000792	AE000790	[3]
Bol26	<i>B. burgdorferi</i> s.s.	Italy	<i>I. ricinus</i>	ABCW02000001-4	CP001568	CP001571	[2]
CA-11.2A	<i>B. burgdorferi</i> s.s.	USA (California)	<i>I. pacificus</i>	ABJY02000001-14	CP001484	CP001473	[2]
JD1	<i>B. burgdorferi</i> s.s.	USA (Massachusetts)	<i>I. scapularis</i>	CP002312	CP002316	CP001652	[2]
N40	<i>B. burgdorferi</i> s.s.	USA (New York)	<i>I. scapularis</i>	CP002228	CP002239	CP001651	[2]
WI91-23	<i>B. burgdorferi</i> s.s.	USA (Wisconsin)	Bird	-	CP001446	CP001447	[2]
ZS7	<i>B. burgdorferi</i> s.s.	Germany	<i>I. ricinus</i>	CP001205	CP001212	CP001199	[2]
118a	<i>B. burgdorferi</i> s.s.	USA (New York)	Human	ABGI02000001-8	CP001535	CP001542	[2]
156a	<i>B. burgdorferi</i> s.s.	USA (New York)	Human	ABCV02000001	CP001271	CP001257	[2]
297	<i>B. burgdorferi</i> s.s.	USA (Connecticut)	Human	-	CP002268	CP001653	[2]
29805	<i>B. burgdorferi</i> s.s.	USA (Connecticut)	<i>I. scapularis</i>	ABJX02000001-38	CP001550	CP001554	[2]
64b	<i>B. burgdorferi</i> s.s.	USA (New York)	Human	ABKA02000001-6	CP001422	CP001421	[2]

Strain	Genospecies	Geographical origin	Biological source	Accession Number Chr	Accession Number cp26	Accession Number lp54	Reference(s)
72a	<i>B. burgdorferi</i> s.s.	USA (New York)	Human	ABGJ02000001-6	CP001375	CP001370	[2]
PKo	<i>B. afzelii</i>	Germany	Human	CP002933	CP002934	CP002950	[1]
ACA-1	<i>B. afzelii</i>	Sweden	Human	ABCU02000001-2	CP001250	CP001247	[1]
PBi	<i>B. bavariensis</i>	Germany	Human	CP000013	NC_006128	CP000015	[4]
SV1	<i>B. finlandensis</i>	Finland	<i>I. ricinus</i>	ABJZ02000001-5	CP001522	CP001524	[5]
VS116	<i>B. valaisiana</i>	Switzerland	<i>I. ricinus</i>	NZ_ABCY02000001	CP001432	CP001433	[6]
A14S	<i>B. spielmanii</i>	The Netherlands	<i>I. ricinus</i>	ABKB02000001-13	CP001467	CP001469	[6]
DN127	<i>B. bissettii</i>	California	<i>I. pacificus</i>	CP002746	CP002747	CP002761	[6]
<ol style="list-style-type: none"> 1. Casjens SR, Mongodin EF, Qiu W-G, Dunn JJ, Luft BJ, et al. (2011) Whole-genome sequences of two <i>Borrelia afzelii</i> and two <i>Borrelia garinii</i> Lyme disease agent isolates. J Bacteriol 193: 6995–6996. 2. Schutzer SE, Fraser-Liggett CM, Casjens SR, Qiu W-G, Dunn JJ, et al. (2011) Whole-genome sequences of thirteen isolates of <i>Borrelia burgdorferi</i>. J Bacteriol 193: 1018–1020. 3. Fraser C, Casjens S, Huang W, Sutton GG, Clayton R, et al. (1997) Genomic sequence of a Lyme disease spirochaete, <i>Borrelia burgdorferi</i>. Nature 390. 4. Glöckner G, Lehmann R, Romualdi A, Pradella S, Schulte-Spechtel U, et al. (2004) Comparative analysis of the <i>Borrelia garinii</i> genome. Nucleic Acids Res 32: 6038–6046. 5. Casjens SR, Fraser-Liggett CM, Mongodin EF, Qiu W-G, Dunn JJ, et al. (2011) Whole genome sequence of an unusual <i>Borrelia burgdorferi</i> sensu lato isolate. J Bacteriol 193: 1489–1490. 6. Schutzer SE, Fraser-Liggett CM, Qiu W-G, Kraiczy P, Mongodin EF, et al. (2012) Whole-genome sequences of <i>Borrelia bissettii</i>, <i>Borrelia valaisiana</i>, and <i>Borrelia spielmanii</i>. J Bacteriol 194: 545–546 							

Table S4

	M0	M1	M2	M3	M4	M5
M0	-	0.88	0.54	1.24	0.83	0.006
M1	1.14	-	0.61	1.40	0.94	0.007
M2	1.86	1.63	-	2.29	1.54	0.011
M3	0.81	0.71	0.43	-	0.67	0.005
M4	1.21	1.06	0.65	1.49	-	0.007
M5	167	147	90	206.5	138.5	-

CHAPITRE 2

DIFFERENCIATION HÔTE-SPECIFIQUE ENTRE LES POPULATIONS DE *B. BURGDORFERI* S.L.

Résumé de l'article 2

En Europe, les différentes génoespèces de *B. burgdorferi* s.l. sont connues pour être spécifiques des mammifères, des oiseaux ou des reptiles [Kurtenbach *et al.*, 2006 ; Margos *et al.*, 2011]. En revanche, peu d'études ont décrit, à une échelle plus fine, l'association entre les différents génotypes de *B. burgdorferi* s.l. et les espèces hôtes. Dans ce contexte épidémiologique, ce chapitre est dédié à l'étude de la diversité des génotypes qui infectent différentes espèces de mammifères et à leur caractérisation dans le but d'identifier les différents réservoirs de la bactérie. Entre 2005 et 2010, des rongeurs de trois espèces ont été échantillonnés dans une forêt péri-urbaine près de Paris (forêt de Sénart) grâce à une collaboration entre membres du Muséum National d'Histoire Naturelle, de l'Institut Pasteur et de l'unité d'épidémiologie Animale. Les différents génotypes de *B. burgdorferi* s.l. infectant environ 230 rongeurs ont été caractérisés par le séquençage 454 des produits PCR du gène de ménage *rplB* et du gène *ospC* qui ne semblent, comme décrit dans l'Article 1, pas soumis aux mêmes contraintes évolutives. Avec l'aide de membres de l'unité d'Épidémiologie Animale, j'ai réalisé les analyses dont les résultats ont été valorisés par un article publié dans la revue *PLoS ONE* au début de l'année 2014.

Avantages du séquençage haut-débit et études des co-infections

Dans cette étude, l'utilisation du séquençage haut-débit a non seulement permis d'établir les relations de parentés entre les différents génotypes de *B. burgdorferi* s.l. comme il aurait été possible de le faire avec une approche de MLST. Cependant, elle a également autorisé l'étude des co-infections des individus par différents génotypes avec une résolution plus fine que ce que les approches, classiquement utilisés pour aborder ce type de problème, ne l'auraient permis *via* notamment la détection de groupes génétiques sub-spécifiques.

Différenciation en fonctions des hôtes pour les deux marqueurs

Les analyses phylogénétiques et de variance moléculaire ont montré que pour les deux génoespèces de *Borrelia* majoritairement retrouvées, *B. burgdorferi* s.s. et *B. afzelii*, les populations associées aux différentes espèces d'hôtes échantillonnés sont significativement différenciées et forment des ensembles génétiques distincts et ce, pour chacun des deux marqueurs séquencés.

Chapitre 2

*Association entre les génotypes de *rplB* et d'*ospC* et patrons de co-infections*

L'étude des co-occurrences des génotypes des deux marqueurs chez les individus hôtes ont montré que plusieurs génotypes d'*ospC* sont associés à chaque génotype de *rplB* probablement à cause des fréquentes recombinaisons qui impactent le gène *ospC*. Cette analyse a également permis de montrer que les différents hôtes ne sont pas co-infectés avec la même fréquence. Le tamia de Sibérie, qui est porteur d'une plus grande diversité de génotypes que le campagnol roussâtre, semble être un hôte particulièrement sujet aux co-infections. Ceci peut s'expliquer par plusieurs hypothèses non exclusives. Tout d'abord puisque c'est une espèce qui a été introduite récemment dans la forêt étudiée, elle est peut-être plus sensible aux infections par *B. burgdorferi* s.l.. Ensuite, la probabilité de contact du tamia avec la bactérie est sans doute plus importante que pour les autres espèces étudiées puisqu'elle a une charge entomologique plus importante. Enfin, plus de génotypes bactériens sont peut être capables d'infecter cette espèce.

ARTICLE 2

High-Throughput sequence typing reveals genetic differentiation and host specialization among populations of the *Borrelia* *burgdorferi* species complex that infect rodents

Maude Jacquot, Maxime Bisseux, David Abrial, Maud Marsot,
Elisabeth Ferquel, Jean-Louis Chapuis, Gwenaël Vourc'h, Xavier
Bailly

High-Throughput Sequence Typing Reveals Genetic Differentiation and Host Specialization among Populations of the *Borrelia burgdorferi* Species Complex that Infect Rodents

Maude Jacquot^{1*}, Maxime Bisseux^{1a}, David Abrial¹, Maud Marsot^{1,2ab}, Elisabeth Ferquel^{3ac}, Jean-Louis Chapuis², Gwenaél Vourc'h¹, Xavier Bailly¹

1 INRA, UR346 Épidémiologie Animale, Saint Genès Champanelle, France, **2** MNHN-CNRS-P6, UMR 7204 Conservation des Espèces Restauration et Suivi des Populations, Paris, France, **3** Institut Pasteur, CNR *Borrelia*, Paris, France

Abstract

Lyme disease is a zoonosis caused by various species belonging to the *Borrelia burgdorferi* bacterial species complex. These pathogens are transmitted by ticks and infect multiple, taxonomically distinct, host species. From an epidemiological perspective, it is important to determine whether genetic variants within the species complex are able to spread freely through the whole host community or, instead, if certain variants are restricted to particular hosts. To this end, we characterized the genotypes of members of the *B. burgdorferi* species complex; the bacteria were isolated from more than two hundred individuals captured in the wild and belonging to three different rodent host species. For each individual, we used a high-throughput approach to amplify and sequence *rplB*, a housekeeping gene, and *ospC*, which is involved in infection. This approach allowed us to evaluate the genetic diversity both within and among species in the *B. burgdorferi* species complex. Strong evidence of genetic differentiation among host species was revealed by both genes, even though they are, *a priori*, not constrained by the same selective pressures. These data are discussed in the context of the advancements made possible by multi-locus high-throughput sequencing and current knowledge of Lyme disease epidemiology.

Citation: Jacquot M, Bisseux M, Abrial D, Marsot M, Ferquel E, et al. (2014) High-Throughput Sequence Typing Reveals Genetic Differentiation and Host Specialization among Populations of the *Borrelia burgdorferi* Species Complex that Infect Rodents. PLoS ONE 9(2): e88581. doi:10.1371/journal.pone.0088581

Editor: Brian Stevenson, University of Kentucky College of Medicine, United States of America

Received: November 29, 2013; **Accepted:** January 7, 2014; **Published:** February 12, 2014

Copyright: © 2014 Jacquot et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the French National Institute for Agricultural Research (INRA) through the 'AIP Bioressources 2011' program. Maude Jacquot is supported by a PhD fellowship from the French Ministry of Higher Education and Research (MESR). Interactions between members of the INRA and the Pasteur Institute are funded by the 'Tick and Tick-Borne Diseases' (TMT) group, which is part of the Research Consortium on Biological Interactions (GdR REID). Field sampling was conducted with the financial support of the Conseil Régional d'Île-de-France and the Office National des Forêts. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: maude.jacquot@gmail.com

a Current address: Université d'Auvergne, EPIE Laboratoire de Virologie, Clermont-Ferrand, France

b Current address: ANSES, Unité Épidémiologie, Maisons-Alfort, France

c Current address: Institut Pasteur, Département de Virologie, Paris, France

Introduction

A large percentage of the pathogens that infect humans are zoonotic [1] and are maintained in local host reservoirs. Pathogens can be specific to a single host species, or they can possess a more generalist transmission strategy and infect multiple, unrelated host species [2]. Identifying these hosts remains a challenging task that can have crucial implications for infectious disease management strategies [3].

Studying the genetic diversity of zoonotic pathogens can provide valuable insights into various aspects of their epidemiology. For example, analyses of the diversity of pathogens within infected individuals can reveal information about the exposure and/or the susceptibility of hosts to different pathogen genotypes. Furthermore, cases of host co-infection, in which a host is simultaneously infected by different pathogen genotypes, are influenced by the contact dynamics between hosts and infectious

agents, the establishment of host defenses, and potential interactions between and among pathogen genotypes. Finally, observations of genetic differentiation between pathogens that infect different host species provide information about genetic exchanges between pathogens found in different potential reservoirs and potential specialization events.

With these goals in mind, the *Borrelia burgdorferi* species complex is an ideal model system in which to study the spread of zoonotic pathogens within a community of hosts. This species complex includes the pathogens that cause Lyme borreliosis, one of the most frequent vector-borne zoonotic diseases in the Northern Hemisphere. Bacteria of the *B. burgdorferi* species complex are acquired and transmitted by ticks during their blood meals and infect many different host species, including small mammals, birds, and reptiles. While species of the complex are thought to have host ranges restricted to one or a few host species, the host specificity of individual bacterial genotypes has yet to be evaluated [4]. Finally,

this species complex also provides an opportunity to study co-infection, as individual hosts are usually infested by several ticks, each of which can transmit a different bacterial genotype [5].

The process of infection by members of the *B. burgdorferi* species complex involves many different bacterial genes. Among these, the outer surface protein encoded by the *ospC* gene is characterized by a peculiar pattern of polymorphism, one that is shaped by diversifying selective pressures that arise as a function of host diversity at both a genetic and an immunological scale. In addition, recombination plays an important role in the evolution of this genome region [6], and the shuffling of genotypes caused by recombination could result in the disruption of any potential associations between *ospC* genotypes and those of housekeeping genes. However, residual association might still result from either selective or epidemiological constraints that induce linkage disequilibrium at the whole genome scale. Therefore, by comparing the patterns of diversity found in housekeeping genes with those found in infection-related genes, we can obtain information about the evolution of the *B. burgdorferi* species complex.

Recent developments in sequencing technology have provided the opportunity to investigate the genetic diversity of pathogens within and among hosts in greater detail. Massive parallel sequencing can be used to obtain numerous sequences from the pathogens within a single host and to characterize multiple loci within each pathogen. Moreover, pathogens infecting several hosts can be characterized in a single high-throughput sequencing run, providing an efficient way to perform studies across a community of hosts.

In this context, the aim of this study was to describe patterns of genetic differentiation within and among populations of the *B. burgdorferi* species complex that infect three rodent species. Hosts were sampled in a French peri-urban forest over the course of several years, and bacteria were characterized using High-Throughput Multi-Locus Sequence Typing (HiMLST) [7]. This aimed at obtaining an accurate description of genotypes that infect the studied hosts. Characterization of pathogens targeted both a housekeeping gene and an infection-related gene, differentiated by different phylogenies, to evaluate potential evolutionary scenarios. Results are discussed in the context of current knowledge of molecular epidemiology into the *B. burgdorferi* species complex.

Materials and Methods

Ethic statement

All conducted experiments complied with the current laws of France. Trapping and collection of rodents conducted on the study site (Forêt de Sénart, Essonne, France) were carried out under the control of Laurent TILLON (Office National des Forêts), Head of Research Group mammals. The project was approved by the Ethics Committee in Animal Experiment (CEMEA Auvergne). Rodents were euthanized by cervical dislocation. Ear punch biopsies were limited to the minimum size needed, the puncture was disinfected with hydrogen peroxide and checked before releasing the animal.

Sample selection

As part of another study, Siberian chipmunks (*Tamias sibiricus barberi*), bank voles (*Myodes glareolus*), and wood mice (*Apodemus sylvaticus*) were caught during field sampling performed in the forest of Sénart (3200 ha, 48°40'N, 02°29'E), located near Paris (France), over the course of six years, from 2005 to 2010. An ear biopsy was taken from each animal, from which we extracted DNA; PCR Restriction Fragment Length Polymorphism (RFLP)

analysis was then used to test whether the animals were infected with bacteria belonging to the *B. burgdorferi* species complex [8] [9].

As the aim of this study was to evaluate, in depth, the genotypic diversity present within these hosts, we selected 228 infected individuals (Table S1) using a random stratified sampling strategy, targeting a maximum of 30 individuals per host species per year, as long as there were enough samples; the sample contained 125 chipmunks, 93 bank voles, and 10 wood mice.

Sample characterization

The purpose of this study was to compare the patterns of genetic differentiation observed in a housekeeping gene with those from a gene obviously affected by host-driven selective pressures. From among the different markers included in the MLST scheme that has been developed for the *B. burgdorferi* species complex, we focused on partial sequences of *rplB*, a housekeeping gene. We also studied partial sequences of *ospC*, an infection-related gene with antigenic properties that is affected by balancing and/or negative frequency-dependent selective pressures that arise from host-pathogen interactions [10] [5]. Evidence of these evolutionary patterns can be observed in *ospC* sequence data and seem to be widespread among species within the *B. burgdorferi* species complex.

Markers were amplified independently via semi-nested PCR for each DNA sample using the GoTaq kit (Promega, Fitchburg, USA). Outer amplifications were performed using the modified primers *rplB_Out_F* (5' AGGGTATTAAGACTTATAAGC 3') and *rplB_Out_R* (5' AGGCTGTCCCCAAGGAGAYAC 3') for *rplB* [11]. For *ospC*, primers *ospC_F1* (5' GGGAWCCAAAATCTAATAYAA 3') and *ospC_R1* (5' ATATTGACTTTATTTTCCAGTTAC 3') were used [6]. Amplifications were performed using a total volume of 25 µl, which included 8 µl of DNA extract, 0.25 µl of Taq polymerase, 10 pmol of each primer, and the buffer provided by the manufacturer at a final concentration of 1X. Amplifications were performed using the following reaction cycle parameters: an initial denaturation step at 95°C for 10 min, followed by 40 cycles, each composed of a denaturation step at 94°C for 30 s, an annealing step of 30 s (at 58°C for *rplB* or 56°C for *ospC*), and an elongation step at 72°C for 40 s. These cycles were followed by a final elongation step of 5 min at 72°C.

For the inner PCR amplifications, the forward primers for each region were composed of two parts: 1) a 30-base-pair (bp) oligonucleotide required for the unidirectional sequencing process on a 454 platform (5' CCTATCCCCTGTGTGCCCTGGCAGTCTCAG 3') and 2) a gene-specific oligonucleotide to bind the sequence region of interest. The target-specific oligonucleotides were the primers *rplB_In_F* (5' CGCTATAAGACGACTT-TATC 3') for *rplB* [11] and *ospC_F2* (5' AAAAGGAGGCA-CAAATTAATG 3') for *ospC* [12]. Similarly, the reverse primer was composed of three parts: 1) a 30-bp oligonucleotide required for sequencing on a 454 platform (5' CCATCTCATCCCTGCGTGTCTCCGACTCAG 3'); 2) a 12-bp oligonucleotide tag corresponding to the individual host from which the DNA had been extracted; and 3) *rplB_Out_R* or *ospC_R1*, which had already been used for the outer PCR amplifications. Inner amplifications were performed using the same thermal cycling parameters as described above, except that the annealing temperature was fixed at 58°C for both sets of primers.

For each locus, 10 µl of the PCR product obtained from the inner amplification reaction was mixed in a 15-ml tube (BD Biosciences, San Jose, USA). The mixture was vortexed, and 2 ml were pipetted and purified using the NucleoSpin PCR Clean-up Kit (Macherey-Nagel, Düren, Germany). The quantity of DNA after purification was assessed by spectrophotometry at 260 and

280 nm (Nanodrop, Thermo Scientific, Wilmington, USA). Equal amounts of purified *rpIB* and *ospC* PCR products were then mixed with each other and sent to GATC-Biotech (Konstanz, Germany) for sequencing on 1/16th of a GS FLX+ run.

Sequence analysis

Assignment of sequences to individuals and loci. Raw sequences were assigned to their respective individual host animals and target loci using a BLAST approach [13]. We matched the individual-host-specific 12-bp tag at the beginning of each sequence with our reference tag sequences using BLASTN. To ensure the best possible assignment for each sequence, BLAST searches were performed using the following parameters: the query strands to search against database for BLASTN was fixed at 1; the reward for nucleotide match was fixed at 4; the penalty for a nucleotide mismatch was fixed at 5; the gap opening cost was fixed at 3; the gap extending cost was fixed at 5; and the used word size was fixed at 4. The best hit was saved and used to assign each sequence to an individual host.

To assign each sequence to its respective locus (*rpIB* or *ospC*), we then used BLASTN to compare each sequence to those in a database that contained different representative sequences of the target loci from the *B. burgdorferi* MLST database (<http://borrelia.mlst.net>) [11] and from GenBank [14]. BLAST searches were performed using the default parameters. Sequences were assigned to a given locus based on the best BLAST hit if the length of the hit was longer than 130 bp with a percentage of identity higher than 80%. Nucleotides corresponding to primers were removed, and assigned sequences were saved for further analysis.

Trimming of the sequence dataset. After checking the distribution of mismatches along the alignments of the assigned sequences, we decided to use only the first 500 bp of the sequences for subsequent analyses (data not shown). We also removed from the dataset all sequences shorter than 350 bp, of which there were relatively few. We constructed two alignments of the respective sequences of each locus of each individual using K-Align [15].

Genotype delineation. A major drawback of high-throughput sequencing methods is the relatively high base-calling error rate. This issue could present a problem for within-species studies of diversity as analyses are likely to be affected by the introduction of low-frequency variants that are, in reality, artifacts of sequencing error. In order to circumvent this problem, genotypes and genotype groups were delineated according to two successive distance-based nearest-neighbor classifications.

First, we used the sequence alignments obtained for each locus from each host individual to define a subset of nucleotide sites that contained only sites with Single-Nucleotide Polymorphisms (SNPs); in order to be included in the subset, each site had to be successfully base-called in more than 40% of sequences and had to contain at least two alleles that were present in more than 5% of the sequences in the alignment. We then computed pairwise distances between the sequences contained in each alignment based on the number of selected SNPs in which they differed. Gapped positions were ignored during pairwise comparisons. Finally, sequences were clustered so that the pairwise distance between a given sequence and another of the same group (i.e. genotype) was equal to or lower than a locus-specific threshold. These thresholds were fixed according to the diversity pattern of each marker. As *rpIB* appeared to contain little intra-specific diversity, we set the threshold for divergence at one nucleotide site. However, we observed more genetic divergence within the *ospC* sequences and therefore set the *ospC* threshold at two sites. For all

of the genotypes identified, we then obtained consensus sequences based on the majority rule for each locus in each individual host.

Next, to identify genotypes that were present in different host individuals, all consensus sequences and singletons obtained from our intra-individual analysis were grouped together according to the algorithm described above. Then, a majority-rule consensus sequence was found for each genotype that was shared across multiple hosts. To ensure robust data, genotypes that were represented by three or fewer sequences in the larger dataset were removed before subsequent analyses.

Differentiation among host species and sampling years. In order to assess the influence of various structuring factors on the observed genetic diversity, we performed an analysis of molecular variance (AMOVA) on the sequence dataset [16], with host species, sampling year, and host individual as structuring factors. This analysis was based on the genotypes defined above, and, within individual hosts, each genotype was weighted according to the number of sequences it contained. In order to prevent bias in the analysis resulting from the impact of inter-specific divergence among *Borrelia* species, genotypes were considered to be genetically equidistant from each other. The analysis was performed with the *amova* function in the *ade4* library in R [17].

Genotype phylogeny and delineation of genotype groups. We used a phylogenetic approach to further examine the diversity of the observed genotypes. For each locus, consensus sequences and reference sequences were aligned to each other using K-Align [15] these reference sequences were the same as those that had been previously used for locus assignments. Phylogenetic searches were performed with a maximum likelihood approach using PhyML [18], and we chose the most appropriate model of evolution for each alignment according to the Akaike Information Criterion (AIC) [19] calculated using the APE library in R [20]. A phylogenetic network was obtained for each locus using the Neighbor-Net method [21] in SPLITSTREE4 [22]. The distance matrices used to create the networks were computed from alignments with PAUP* 4.0 b10 [23] using the best model for each locus, i.e. GTR+G for *rpIB* and GTR+I+G for *ospC* [24] [25]. The substitution rate matrices were estimated via maximum likelihood and assuming empirical nucleotide frequencies. As calculated by the PhyML analysis, the shape parameter of the gamma distribution for *rpIB* was fixed at 0.227; for *ospC*, the proportion of invariable sites and the shape parameter of the gamma distribution were fixed at 1.117 and 0.169, respectively.

Using the networks, we were able to empirically delineate groups of closely related genotypes. We were then able to calculate the proportion of individuals within each host species that was infected by any given group of closely related genotypes.

Genotype associations and co-occurrence. To search for multi-locus associations, i.e. associations between the different genotypes of the two markers, a graph analysis illustrating the frequency of co-occurrence of the different genotype groups within hosts was performed for both marker. First, we constructed an incidence matrix ($m \times n$), with m as the number of individual hosts and n as the number of genotype groups. This matrix described the presence or absence of the different genotype groups among host individuals. We then created a co-occurrence matrix ($n \times n$) that described the amount of co-occurrence of genotype groups within individuals by multiplying the incidence matrix by its transpose. One-half of the resulting matrix provided the information required to build a graph that described the co-occurrence of genotypes over all individual hosts. Second, we evaluated preferential associations among genotype groups using a "greedy" approach [26] to assess modularity. Measures of modularity aim to

determine the adequacy of different classification schemes in representing clusters and divisions in datasets; here, the clusters represented the co-occurrence of genotypes in individual hosts. In the first step of this analysis, each genotype group was considered to be a set. At each step, the algorithm combined two sets such that the resulting classification had the highest degree of modularity as compared to all possible combinations of sets; this was performed until a single set remained. We calculated estimates of modularity for each classification by comparing the fraction of co-occurrence of genotype groups that fall within the sets of the classification with the expected value of the same quantity if the co-occurrences happened randomly without regard for sets of classification [27]. In the final step of the analysis, the classification with the highest modularity from all the generated classifications was selected.

This analysis was performed using the igraph library in R [28]. The co-occurrence matrix was obtained from the incidence matrix using the graph.incidence and the bipartite.projection functions. The classification analysis was performed using the fastgreedy.community function, and graphical output was produced using the tplot function.

Finally, to obtain quantitative information on the frequency of co-infection by the different genotypes identified in the host community, we used the data from each marker to calculate how frequently an individual host was infected by at least two different genotypes.

Results

In total, we obtained 16,913 sequences. Of these, 16,765 of them were assigned to individual hosts and deposited to the Sequence Read Archive (SRA) database under the accession SRP032755.

16,222 sequences, i.e. 95.9% of the total raw sequences, were assigned to individual hosts and one of the two loci. In the sample of all of the 228 studied mammals, we obtained, on average, 33.2 *rplB* sequences and 37.9 *ospC* sequences per host individual. A few individuals did not yield any sequences, but we were able to successfully obtain *rplB* sequences from 96 chipmunks, 92 bank voles, and 8 wood mice, and *ospC* sequences from 94 chipmunks, 90 bank voles, and 5 wood mice.

Using the nearest-neighbor classification algorithm, genotypes were delineated from the sequences for each locus; genotypes containing fewer than four sequences were excluded from subsequent analyses. Consequently, our *rplB* dataset included 33 unique genotypes identified from 87 chipmunks, 87 bank voles, and 7 wood mice. In the phylogenetic analysis based on the *rplB* data, these genotypes formed five major genotype groups (Figure 1). Most of these groups, with the exception of the group G3, were closely related to previously described, reference sequences. Genotype groups G3 and G4, whose members had been recovered from 86 bank voles and 48 chipmunks, respectively, clustered with sequences of *B. afzelii*. Sequences of the G1 genotype group were isolated from 46 chipmunks; the sequences seemed to correspond to those from *B. burgdorferi*, while G2 genotype sequences grouped with the sequence of the *B. spielmanii* *rplB*72 allele. Genotype group G5, whose sequences clustered with those of *B. garinii*, was recovered from only two individual hosts. Within each genotype group, most sequences were obtained from only a single host species, as most members of the G1, G2, and G4 genotypes were isolated from chipmunks and the majority of the G3 sequences came from bank voles. The relatively few sequences recovered from wood mice were found in groups G2 and G4.

In the *ospC* analysis, 80 genotypes were retained after grouping via the nearest-neighbor classification algorithm. The genetic diversity displayed in the phylogenetic network of *ospC* sequences,

when compared with that of the *rplB* network, was consistent with the higher number of genotypes (Figure 2). Based on this network, *ospC* genotypes were delineated into 14 groups, in which, as in the *rplB* data, there was a clear influence of host specificity. Groups G1, G2, G4, G9, G10, G11, G12, G13, and G14 included mainly sequences from chipmunks, while groups G3 and G8 consisted mainly of sequences recovered from bank voles. Sequences isolated from wood mice clustered in genotypes G2 and G10.

We studied the genetic differentiation among *Borrelia* genotypes detected in our sample using an AMOVA, in which the explanatory factors were host species, sampling year, and host individual. As expected from the phylogenetic networks, a substantial percentage of the molecular variance in our sample was explained by the different host species for both loci, 58% for *rplB* and 21% for *ospC* (Tables 1 and 2). Sampling year failed to explain much molecular variance (<1%). Finally, 11 and 17% of the observed variance could be explained by intra-individual variation in *rplB* and *ospC*, respectively.

^adegrees of freedom; ^bsums of squares; ^cmean squares; ^dcomponents of covariance; ^epercentage contribution to the total covariance; ^fPhi statistic.

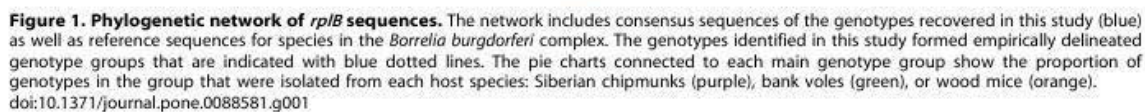
The graph of genotype associations illustrates the frequency of co-occurrence of the different genotype groups for both markers in host individuals (Figure 3). By using measures of modularity to classify groups of co-occurring sequences, we found that the genotype groups isolated from bank voles were distinct from those isolated from chipmunks. We also found evidence that supports the delineation of various species in the *Borrelia burgdorferi* species complex. Among the various associations detected by this analysis, we observed that several *ospC* genotype groups were associated with a single *rplB* genotype. In addition, chipmunks were infected by several *rplB* genotype groups, indicating co-infection by different species of *Borrelia*. Of the individuals from which sequences were recovered, 26% of chipmunks and 11% of bank voles were co-infected by different *rplB* genotypes, while more than 30% of chipmunks, bank voles, and wood mice were co-infected by multiple *ospC* genotypes (Table 3).

Discussion

In previous studies, genotypes of the *B. burgdorferi* species complex have been characterized using multiple techniques, including MLST using Sanger sequencing [29] [30] [31] [11] [32] and PCR-RFLP [33]. These two methods are complementary as the first gives an accurate estimation of the relatedness between bacterial genotypes and the second is more useful in detecting co-infection. Here, we used HiMLST [7] to efficiently characterize the genetic diversity of the *B. burgdorferi* species complex within three sympatric mammal species. As expected, high-throughput sequencing is potentially quite valuable in these types of studies because it combines the advantages of both MLST and PCR-RFLP.

Benefits associated with high-throughput sequence typing

High-throughput sequencing of the *rplB* and the *ospC* markers of the *B. burgdorferi* species complex revealed a high degree of bacterial diversity within infected chipmunks, bank voles, and wood mice in the forest of Sénart (France). The observed *rplB* genotypes were analyzed in the context of pre-existing MLST data and were found to correspond to four species of *Borrelia*: i) *B. afzelii*, including a genotype group that was not described in the dedicated MLST database, ii) *B. burgdorferi* s.s., iii) *B. garinii*, and iv) *B. spielmanii*. When we used this method to assign a species label



distinct genotypes of a single locus, as we observed in the *rplB* house-keeping gene, can provide information about the co-infection of host individuals by different bacterial lineages. Statistical developments that allow researchers to quantify the relative impact of both kinds of co-occurrence patterns will be invaluable in future analyses of similar metagenomic data.

Patterns of association among *ospC* and *rplB* genotypes

As described above, the co-infection of individual hosts in our samples by different genotypes was not rare. For example, our graph analysis revealed the frequent presence in bank voles of two *ospC* variants genetically linked with a single *rpIB* allele. In chipmunks, however, we observed more co-infection events, which can involve multiple combinations of different *Borrelia* species, a finding that could be explained by several hypotheses. First, at these study sites, chipmunks have a higher tick burden than bank voles do [34]. This could result in a higher probability of contact with bacteria of the *B. burgdorferi* species complex and, consequently, a higher risk of co-infection. Second, the chipmunk is considered an invasive species in the Sénart forest; it was introduced there in the early 1970s [35], and the species had colonized the study site by the late 1990s. An increase in the allocation of an individual's resources to reproduction at the cost of allocation to immune responses has been presented as a potential strategy of successful invasive species [36]. It might thus

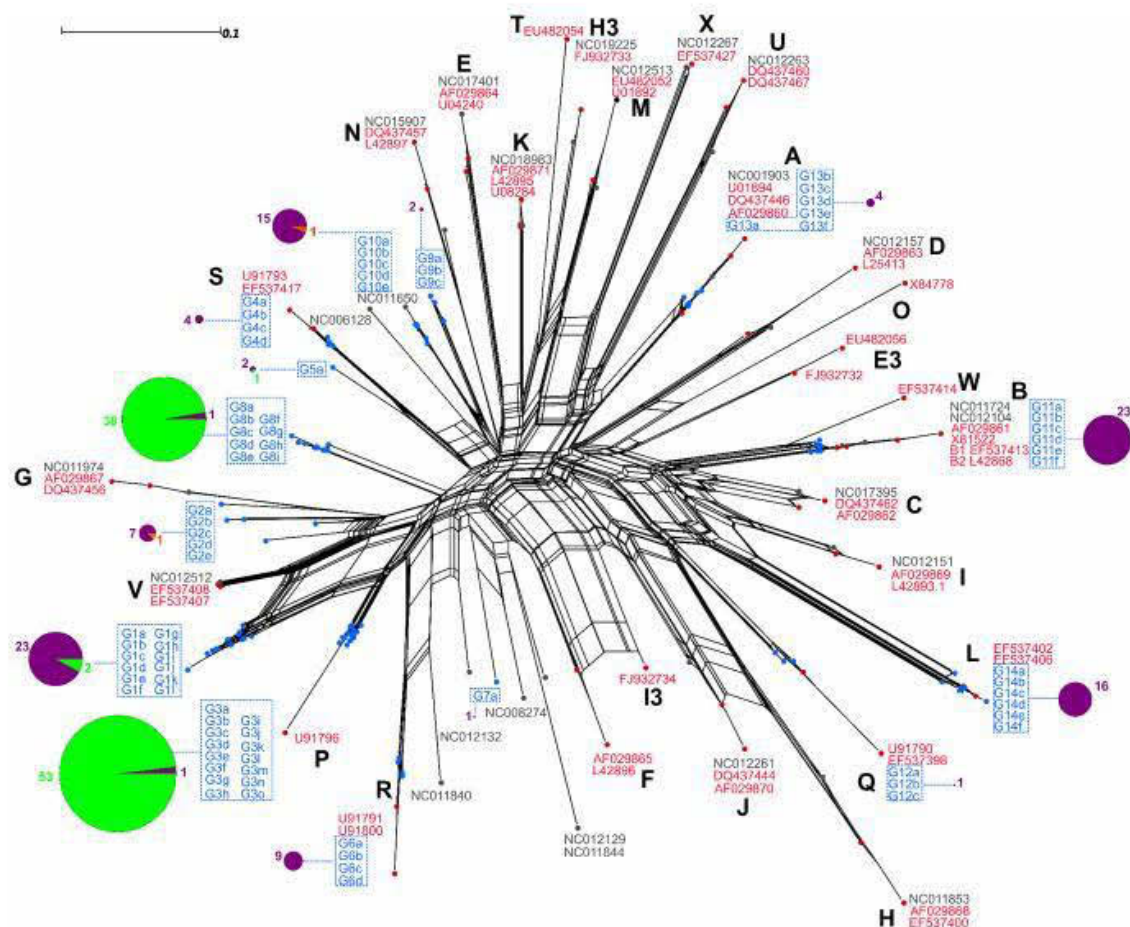


Figure 2. Phylogenetic network of *ospC* sequences. The network includes consensus sequences of the genotypes recovered in this study (blue) as well as reference sequences for species in the *Borrelia burgdorferi* species complex. Genotype groups are delineated by blue dotted lines, and capital letters to the side of each group represent unique *ospC* alleles detected in previous studies [37]. The pie charts connected to each main genotype group show the proportion of genotypes in the group that were isolated from each host species: Siberian chipmunks (purple), bank voles (green), or wood mice (orange).

doi:10.1371/journal.pone.0088581.g002

be that the chipmunk population established in Sénart has a low investment in immune function, resulting in a higher degree of susceptibility to infection by bacteria of the *B. burgdorferi* species

Table 1. Results of the AMOVA performed on *rpIB* sequence data.

Levels of variability	Df ^a	SSq ^b	MSq ^c	S ^d	% ^e	Phi ^f
Among host species	2	1039.14	519.57	0.27	58.19	0.58
Among years within host species	12	77.05	6.42	0.003	0.58	0.01
Among samples within years	172	922.79	5.37	0.14	29.83	0.72
Within samples	7072	373.20	0.053	0.053	11.41	0.89

^adegrees of freedom; ^bsums of squares; ^cmean squares; ^dcomponents of covariance; ^epercentage contribution to the total covariance; ^fPhi statistic.

doi:10.1371/journal.pone.0088581.t001

Table 2. Results of the AMOVA performed on *ospC* sequence data.

Levels of variability	Df ^a	SSq ^b	MSq ^c	S ^d	% ^e	Phi ^f
Among host species	2	439.44	219.72	0.11	20.93	0.21
Among years within host species	11	190.96	17.36	0.001	0.23	0.003
Among samples within years	163	2451.03	15.04	0.32	61.94	0.79
Within samples	8238	713.59	0.087	0.087	16.90	0.83

doi:10.1371/journal.pone.0088581.t002

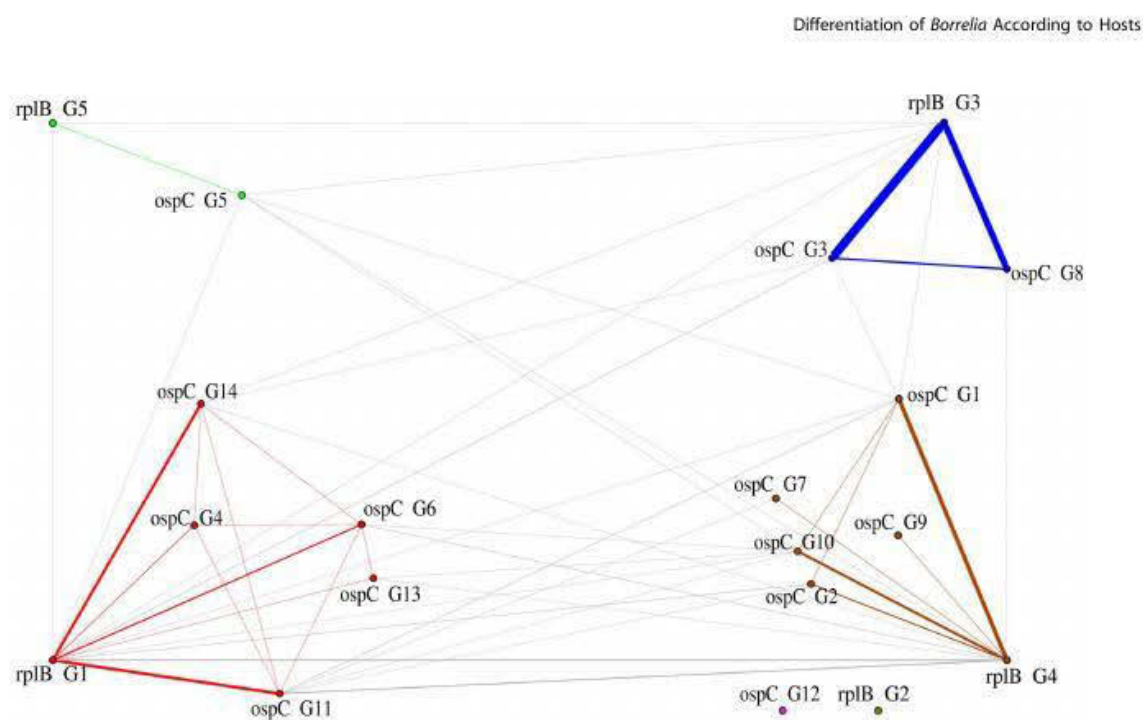


Figure 3. Visual representation of patterns of genotype group associations found in this study. The graph includes genotypes of both target loci and from all three host species. Each distinct community of co-occurring genotype groups is highlighted in a different color, and the thickness of lines is proportional to the frequency of co-occurrence. Vertices were placed empirically; genotypes located towards the top of the figure correspond to those mainly found in bank voles, while those towards the bottom were associated with chipmunks. The cluster of genotypes at the bottom left corresponds to strains of *B. burgdorferi* s.s. and that on the top and bottom right to *B. afzelii*.
doi:10.1371/journal.pone.0088581.g003

complex. Finally, it is possible that more bacterial genotypes are able to infect chipmunks than bank voles.

Differentiation among hosts

The strong genetic differentiation observed between bacterial genotypes that infected bank voles and those that infected chipmunks is compatible with the hypothesis of host-specificity in bacterial populations. In spite of the difference in polymorphism patterns between *rplB* and *ospC*, genotypes that were isolated from chipmunks and bank voles were significantly different from each other at both markers. Indeed, this study did not identify any bacterial genotype that was present at high frequencies in both host species. Our data support previous results, obtained via PCR-RFLP, that show that in a sample of chipmunks and bank voles, genotypes of *B. burgdorferi* s.s. were found only in chipmunks. However, the use of PCR-RFLP did not allow that previous study

to discriminate between genotypes of *B. afzelii* associated with either chipmunks or bank voles [8]. Together with those of the current study, these results demonstrate that chipmunks and bank voles are hosts of independent genotype groups within the *B. burgdorferi* species complex. The ability of different strains of Lyme borreliosis-causing bacteria to persist in two independent host species could be explained by two non-exclusive hypotheses. First, a given host species might be able to clear infections caused by a specific subset of *Borrelia* genotypes or bacterial lineages could differ in their ability to infect the two host species, which would result in different infection cycles for the different genotypes. However, it cannot be excluded that chipmunks and bank voles are simply exposed to different bacterial genotypes as a result of host-vector interactions, a situation that could result in the creation of different, host-specific transmission cycles. For example, if ticks have a higher probability of parasitizing a given host species, some degree of differentiation in both housekeeping and virulence-related genes would likely result. It would therefore be interesting to compare samples from questing ticks and feeding ticks to determine if there are changes in the diversity of genotypes of the *B. burgdorferi* species complex.

Origin of low frequency infections

Unlike chipmunks and bank voles, wood mice are only infrequently infected by strains of the *B. burgdorferi* species complex [8], and all but one of the genotypes identified from wood mice were shared with the two other host species. This finding could indicate that wood mice are able to become infected but do not play a critical role in the dynamic of the *B. burgdorferi* species complex in Sénart forest. This infection pattern in wood mice

Table 3. Percentage (%) of host individuals of each host species in which co-infection by at least two genotypes of the *Borrelia burgdorferi* species complex was detected.

Host species	<i>rplB</i>	<i>ospC</i>
<i>Tamias sibiricus barberi</i>	19.54	38.64
<i>Myodes glareolus</i>	4.55	31.33
<i>Apodemus sylvaticus</i>	0	0

Results are shown for both the *rplB* and *ospC* datasets.
doi:10.1371/journal.pone.0088581.t003

might be due to either i) a relative paucity of bacterial strains that are adapted to infect wood mice, or ii) the low susceptibility to *Borrelia* of wood mice in Sénart. The infection of a few chipmunks and bank voles by atypical (e.g., *B. garinii*) and/or low frequency genotypes might illustrate similar infection patterns.

Perspectives

To conclude, this study used high-throughput sequencing to reveal that strains of the *Borrelia burgdorferi* species complex that were isolated from three sympatric rodent host species presented strong evidence of genetic differentiation at two different loci subject to distinct evolutionary pressures. This result strengthens the need of experimental studies to investigate properly the adaptation of host specific lineages to assess more accurately the dynamic of Lyme disease bacteria among hosts. The efficiency of the HiMLST procedure opens new avenues of study and offers exciting opportunities for further characterization of the diversity and evolution of pathogens infecting host species and vectors. It is indeed important to take into account the variability of pathogens within samples to i) assess properly their population genetic structure and, ii) compare the diversity among infected individuals and eventually at a tissue scale within individuals.

References

1. Woolhouse MEJ, Gowtage-Sequeria S (2005) Host range and emerging and reemerging pathogens. *Emerg Infect Dis* 11: 1842–1847.
2. Haydon DT, Cleaveland S, Taylor LH, Laurensen MK (2002) Identifying reservoirs of infection: a conceptual and practical challenge. *Emerg Infect Dis* 8: 1468–1473.
3. Keesing F, Belden LK, Daszak P, Dobson A, Harvell CD, et al. (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* 468: 647–652.
4. Margos G, Vollmer S a, Ogden NH, Fish D (2011) Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. *Infect Genet Evol* 11: 1545–1563.
5. Brisson D, Dykhuizen DE (2004) *ospC* diversity in *Borrelia burgdorferi*: different hosts are different niches. *Genetics* 168: 713–722.
6. Barbour AG, Trivinsky B (2010) Evolution and distribution of the *ospC* gene, a transferable serotype determinant of *Borrelia burgdorferi*. *MBio* 1: 1–11.
7. Boers SA, van der Reijden WA, Jansen R (2012) High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS One* 7: e39630.
8. Marsot M, Sigaud M, Chapuis JL, Ferquel E, Cornet M, et al. (2011) Introduced Siberian chipmunks (*Tamias sibiricus barkeri*) harbor more-diverse *Borrelia burgdorferi* sensu lato genospecies than native bank voles (*Myodes glareolus*). *Appl Environ Microbiol* 77: 5716–5721.
9. Marsot M, Chapuis J-L, Gasqui P, Dozières A, Masséglia S, et al. (2013) Introduced Siberian chipmunks (*Tamias sibiricus barkeri*) contribute more to Lyme borreliosis risk than native reservoir rodents. *PLoS One* 8: e55377.
10. Haven J, Vargas LC, Mongodin EF, Xue V, Hernandez Y, et al. (2011) Pervasive Recombination and Sympatric Genome Diversification Driven by Frequency-Dependent Selection in *Borrelia burgdorferi*, the Lyme disease Bacterium. *Genetics* 189: 951–966.
11. Margos G, Gatewood AG, Aanensen DM, Hanincová K, Terekhova D, et al. (2008) MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proc Natl Acad Sci U S A* 105: 8730–8735.
12. Theisen M, Borre M, Mathiesen MJ, Mikkelsen B, Lebech A-M, et al. (1995) Evolution of the *Borrelia burgdorferi* outer surface protein *OspC*. *J Bacteriol* 177: 3036–3044.
13. Altschul S, Gish W, Miller W, Myers AW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol*.
14. Benson D a, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. *Nucleic Acids Res* 41: D36–42.
15. Lassmann T, Sonnhammer ELL (2003) Kaluga - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6: 298.

Supporting Information

Table S1 Information on the host individuals sampled and the identity of the *Borrelia* species that were detected in each sample via PCR-RFLP [8] [9] and/or HiMLST. Species labels were assigned to genotypes in this study based on the genotype groups present in the *rplB* data. Genotype G1 was assigned to *Borrelia burgdorferi* s.s., G3 and G4 to *Borrelia afzelii*, and G5 to *Borrelia garinii*.⁰ means that the species was not detected by PCR-RFLP or HiMLST; ¹that the species was detected only by PCR-RFLP; ²that species was detected only by HiMLST; and ³species was detected by both methods; ⁸no sequences were retained for this individual after genotypes containing fewer than four sequences were removed. (DOC)

Acknowledgments

We thank the staff of the GATC, especially Zélie Dubreucq and Marie-Elise Galoppini, for handling the sequencing process. We also thank Jessica Pearce and Lindsay Higgins for proofreading and editing our manuscript.

Author Contributions

Conceived and designed the experiments: XB. Performed the experiments: MM JLC EF GV XB. Analyzed the data: MJ MB XB. Contributed reagents/materials/analysis tools: MJ MB DA XB. Wrote the paper: MJ GV XB.

16. Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 491: 479–491.
17. Dray S, Dufour A (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *J Stat Softw* 22.
18. Guindon S, Gascuel O (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* 52: 696–704.
19. Akaike H (1974) A new look at the statistical model identification. *Autom Control IEEE Trans* 1974.
20. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20: 289–290.
21. Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* 21: 255–265.
22. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23: 254–267.
23. Swofford D (2003) PAUP*: phylogenetic analysis using parsimony, version 4.0 b10: 2003.
24. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci*.
25. Gu X, Fu Y, Li W (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol*.
26. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 70: 066111.
27. Newman M (2004) Fast algorithm for detecting community structure in networks. *Phys Rev E*: 1–5.
28. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal, Complex Syst*.
29. Qiu W-G, Schutzer SE, Bruno JF, Attie O, Xu Y, et al. (2004) Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing. *Proc Natl Acad Sci U S A* 101: 14150–14155.
30. Bunikis J, Garpino U, Tsao J, Berglund J, Fish D, et al. (2004) Sequence typing reveals extensive strain diversity of the Lyme borreliosis agents *Borrelia burgdorferi* in North America and *Borrelia afzelii* in Europe. *Microbiology* 150: 1741–1755.
31. Richter D, Postic D, Sertour N, Livey I, Matuschka F-R, et al. (2006) Delineation of *Borrelia burgdorferi* sensu lato species by multilocus sequence analysis and confirmation of the delineation of *Borrelia spielmanii* sp. nov. *Int J Syst Evol Microbiol* 56: 873–881.
32. Christova I, Van De Pol J, Yazari S, Velo E, Schouls L (2003) Identification of *Borrelia burgdorferi* sensu lato, *Anaplasma* and *Ehrlichia* species, and spotted fever group Rickettsiae in ticks from Southeastern Europe. *Eur J Clin Microbiol Infect Dis* 22: 535–542.
33. Postic D, Assous M V., Grimont PAD, Baranton G (1994) Diversity of *Borrelia burgdorferi* Sensu Lato Evidenced by Restriction Fragment Length Polymorphism

Differentiation of *Borrelia* According to Hosts

- of rrf (5S)-rrl (23S) Intergenic Spacer Amplicons. Int J Syst Bacteriol 44: 743–752.
34. Pisanu B, Marsot M, Marmet J, Chapuis J-L, Réale D, et al. (2010) Introduced Siberian chipmunks are more heavily infested by ixodid ticks than are native bank voles in a suburban forest in France. Int J Parasitol 40: 1277–1283.
35. Chapuis J, Obolenskaya E, Pisanu B, Lissovsky A (2011) Datasheet on *Taxius sibiricus*. CABI, Wallingford, UK: 2011.
36. Lee KA, Klasing KC (2004) A role for immunology in invasion biology. Trends Ecol Evol 19: 523–529.
37. Rudenko N, Golovchenko M, Hönig V, Mallatová N, Krbková L, et al. (2013) Detection of *Borrelia burgdorferi* sensu stricto *ospC* alleles associated with human lyme borreliosis worldwide in non-human-biting tick *Ixodes affinis* and rodent hosts in Southeastern United States. Appl Environ Microbiol 79: 1444–1453.

Données supplémentaires de l'article 2

Table S1

Individual name	Host species	Sampling year	<i>B. afzelii</i>	<i>B. burgdorferi</i> s.s.	<i>B. garinii</i>
S05Cg006	<i>Myodes glareolus</i>	2005	3	0	0
S05Cg008	<i>Myodes glareolus</i>	2005	3	0	0
S05Cg029a	<i>Myodes glareolus</i>	2005	3	0	0
S05Cg032	<i>Myodes glareolus</i>	2005	3	0	0
S05Cg048	<i>Myodes glareolus</i>	2005	3	0	0
S05Cg051	<i>Myodes glareolus</i>	2005	1	2	0
S05Cg057	<i>Myodes glareolus</i>	2005	3	0	0
S05Cg060	<i>Myodes glareolus</i>	2005	X	X	X
S05Cg062	<i>Myodes glareolus</i>	2005	X	X	X
S06Cg001	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg012	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg017	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg022	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg023	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg025	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg038	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg041	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg043	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg044	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg045	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg048	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg051	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg059	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg060	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg061	<i>Myodes glareolus</i>	2006	3	0	0
S06Cg074	<i>Myodes glareolus</i>	2006	3	0	0
S07Cg003	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg013	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg045	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg055	<i>Myodes glareolus</i>	2007	3	0	0

Chapitre 2

S07Cg090	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg093	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg094	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg97	<i>Myodes glareolus</i>	2007	X	X	X
S07Cg098	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg101	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg105	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg112	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg113	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg117	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg123	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg132	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg149	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg156	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg158	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg164	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg175	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg219	<i>Myodes glareolus</i>	2007	X	X	X
S07Cg221	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg226	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg231	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg247	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg262	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg267	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg279	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg286	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg292	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg295	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg299	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg302	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg311	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg315	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg324	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg330	<i>Myodes glareolus</i>	2007	3	0	0
S07Cg334	<i>Myodes glareolus</i>	2007	3	0	0

S07Cg337	<i>Myodes glareolus</i>	2007	3	0	0
S08Cg101	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg102	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg103	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg104	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg109	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg111	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg113	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg115	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg121	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg132	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg178	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg181	<i>Myodes glareolus</i>	2008	X	X	X
S08Cg196	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg221	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg226	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg235	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg251	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg349	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg383	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg384	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg391	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg395	<i>Myodes glareolus</i>	2008	3	0	0
S08Cg396	<i>Myodes glareolus</i>	2008	3	0	0
S09Cg004	<i>Myodes glareolus</i>	2009	3	0	3
S09Cg006	<i>Myodes glareolus</i>	2009	3	0	0
S10Cg023	<i>Myodes glareolus</i>	2010	3	0	0
S10Cg064	<i>Myodes glareolus</i>	2010	X	X	X
S05Ts002	<i>Tamias sibiricus</i>	2005	3	0	0
S05Ts014	<i>Tamias sibiricus</i>	2005	3	0	0
S05Ts017	<i>Tamias sibiricus</i>	2005	3	0	0
S05Ts018	<i>Tamias sibiricus</i>	2005	1	3	0
S05Ts023	<i>Tamias sibiricus</i>	2005	0	3	0
S05Ts024	<i>Tamias sibiricus</i>	2005	3	2	0
S05Ts032	<i>Tamias sibiricus</i>	2005	3	0	0

Chapitre 2

S05Ts034	<i>Tamias sibiricus</i>	2005	3	0	0
S06Ts001	<i>Tamias sibiricus</i>	2006	X	X	X
S06Ts003	<i>Tamias sibiricus</i>	2006	3	2	0
S06Ts008	<i>Tamias sibiricus</i>	2006	X	X	X
S06Ts010	<i>Tamias sibiricus</i>	2006	X	X	X
S06Ts014	<i>Tamias sibiricus</i>	2006	0	3	0
S06Ts015	<i>Tamias sibiricus</i>	2006	0	3	0
S06Ts016	<i>Tamias sibiricus</i>	2006	X	X	X
S06Ts017	<i>Tamias sibiricus</i>	2006	0	3	0
S06Ts018	<i>Tamias sibiricus</i>	2006	3	0	0
S06Ts026	<i>Tamias sibiricus</i>	2006	X	X	X
S06Ts035	<i>Tamias sibiricus</i>	2006	3	0	0
S06Ts061	<i>Tamias sibiricus</i>	2006	X	X	X
S06Ts063	<i>Tamias sibiricus</i>	2006	X	X	X
S07Ts003	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts004	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts005	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts006	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts009	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts012	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts015	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts019	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts023	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts025	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts027	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts031	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts043	<i>Tamias sibiricus</i>	2007	2	3	0
S07Ts051	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts055	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts056	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts057	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts059	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts061	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts064	<i>Tamias sibiricus</i>	2007	2	3	0
S07Ts066	<i>Tamias sibiricus</i>	2007	3	0	0

S07Ts068	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts069	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts070	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts073	<i>Tamias sibiricus</i>	2007	3	2	1
S07Ts076	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts078	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts079	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts081	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts082	<i>Tamias sibiricus</i>	2007	0	2	3
S07Ts084	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts088	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts089	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts090	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts091	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts092	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts098	<i>Tamias sibiricus</i>	2007	3	0	0
S07Ts099	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts100	<i>Tamias sibiricus</i>	2007	0	3	0
S07Ts103	<i>Tamias sibiricus</i>	2007	X	X	X
S07Ts107	<i>Tamias sibiricus</i>	2007	0	3	0
S08Ts014	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts031	<i>Tamias sibiricus</i>	2008	X	X	X
S08Ts035	<i>Tamias sibiricus</i>	2008	2	2	1
S08Ts043	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts061	<i>Tamias sibiricus</i>	2008	X	X	X
S08Ts066	<i>Tamias sibiricus</i>	2008	0	2	1
S08Ts083	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts088	<i>Tamias sibiricus</i>	2008	X	X	X
S08Ts098	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts100	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts113	<i>Tamias sibiricus</i>	2008	2	1	0
S08Ts120	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts124	<i>Tamias sibiricus</i>	2008	1	0	0
S08Ts125	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts128	<i>Tamias sibiricus</i>	2008	2	1	0

Chapitre 2

S08Ts136	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts139	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts146	<i>Tamias sibiricus</i>	2008	2	1	0
S08Ts148	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts150	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts151	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts156	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts158	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts159	<i>Tamias sibiricus</i>	2008	3	0	0
S08Ts161	<i>Tamias sibiricus</i>	2008	2	1	0
S08Ts166	<i>Tamias sibiricus</i>	2008	3	2	0
S08Ts167	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts176	<i>Tamias sibiricus</i>	2008	2	3	0
S08Ts178	<i>Tamias sibiricus</i>	2008	0	3	0
S08Ts186	<i>Tamias sibiricus</i>	2008	0	3	0
S09Ts060	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts108	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts109	<i>Tamias sibiricus</i>	2009	3	1	0
S09Ts155	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts156	<i>Tamias sibiricus</i>	2009	3	0	0
S09Ts158	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts159	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts160	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts180	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts180	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts187	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts196	<i>Tamias sibiricus</i>	2009	X	X	X
S09Ts204	<i>Tamias sibiricus</i>	2009	3	0	0
S09Ts214	<i>Tamias sibiricus</i>	2009	X	X	X
S10Ts018	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts029	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts039	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts041	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts062	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts065	<i>Tamias sibiricus</i>	2010	X	X	X

S10Ts091	<i>Tamias sibiricus</i>	2010	3	0	0
S10Ts096	<i>Tamias sibiricus</i>	2010	3	0	0
S10Ts113	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts117	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts123	<i>Tamias sibiricus</i>	2010	X	X	X
S10Ts125	<i>Tamias sibiricus</i>	2010	0	3	0
S10Ts145	<i>Tamias sibiricus</i>	2010	2	1	0
S10Ts157	<i>Tamias sibiricus</i>	2010	3	0	1
S10Ts255	<i>Tamias sibiricus</i>	2010	3	0	0
S10Ts309	<i>Tamias sibiricus</i>	2010	3	0	0
S10Ts316	<i>Tamias sibiricus</i>	2010	0	3	0
S10Ts324	<i>Tamias sibiricus</i>	2010	0	3	0
S10Ts333	<i>Tamias sibiricus</i>	2010	0	3	0
S06As006	<i>Apodemus sylvaticus</i>	2006	3	0	0
S06As14	<i>Apodemus sylvaticus</i>	2006	3	0	0
S07As020	<i>Apodemus sylvaticus</i>	2007	3	0	0
S07As034	<i>Apodemus sylvaticus</i>	2007	X	X	X
S07As048	<i>Apodemus sylvaticus</i>	2007	3	0	0
S07As049	<i>Apodemus sylvaticus</i>	2007	3	0	0
S07As073	<i>Apodemus sylvaticus</i>	2007	3	0	0
S08As083	<i>Apodemus sylvaticus</i>	2008	X	X	X
S08As127	<i>Apodemus sylvaticus</i>	2008	3	0	0
S10As034	<i>Apodemus sylvaticus</i>	2010	X	X	X

CHAPITRE 3

CYCLES DE TRANSMISSION DES GENOTYPES DE *B. BURGDORFERI* S.L. ET ASSIGNATION DES ESPÈCES D'HÔTES À CES CYCLES

Résumé de l'article 3

D'un point de vue épidémiologique, il est important de connaître quels sont les variants génétiques de *B. burgdorferi* s.l. qui circulent chez les hôtes mais également chez les tiques. De plus, nous avons vu au cours de l'introduction, qu'il est d'intérêt de déterminer quelles sont les espèces d'hôtes qui participent le plus à la circulation de l'agent pathogène et donc au risque pour l'homme d'être infecté. Ce troisième chapitre de ma thèse s'intéresse donc à la diversité des génotypes portés par les tiques et à leurs potentiels hôtes d'origine. Pour cela, les génotypes des gènes *rplB* et *ospC* de *Borrelia* infectant 488 tiques de la forêt de Sénart ont été caractérisés par séquençage 454 et ont été comparés aux génotypes de *B. burgdorferi* s.l. infectant les rongeurs (**Article 2**). L'article valorisant les résultats obtenus pour cette étude est en préparation et devrait être prochainement soumis dans la revue *Molecular Ecology*.

Diversité génétique des B. burgdorferi s.l. chez les tiques

L'analyse de diversité de génotypes portés par les tiques a révélé qu'elle était plus importante que celle des génotypes portés par les espèces de rongeurs échantillonnées et utilisées pour l'étude précédente du chapitre 2. En effet, la plupart des nymphes étaient infectées par un ou plusieurs génotypes de *B. burgdorferi* s.s., *B. afzelii*, *B. garinii* et *B. valaisiana* et moins fréquemment par des génotypes de *B. spielmanii* ou *B. andersonii*. Ces résultats suggèrent donc que de nombreux génotypes portés par les tiques sont maintenus chez des hôtes qui n'ont pas été échantillonnés.

Co-occurrences des génotypes et cycle de transmission

L'étude des co-occurrences des génotypes des deux marqueurs *rplB* et *ospC* chez les individus tiques et hôtes, suggère qu'il existe au moins quatre cycles de transmission de *B. burgdorferi* s.l. dans la forêt de Sénart, faisant intervenir les génotypes de différentes génoespèces de *Borrelia* et différentes espèces d'hôtes, notamment : *B. burgdorferi* s.s. et les tamias, *B. afzelii* et les tamias, *B. afzelii* et les campagnols ou *B. garinii* qui doit probablement infecter des oiseaux qui n'ont pas été échantillonnés.

Chapitre 3

Contribution des hôtes

Les résultats obtenus à partir d'un modèle statistique nouvellement construit et basé sur la comparaison des fréquences des différents génotypes chez les hôtes et les tiques, suggèrent que la contribution relative du réservoir comprenant le tamia à la diversité des génotypes observée chez les tiques est très importante (>51 %) tandis que le campagnol semble être un hôte dont la contribution au risque acarologique est très faible (<4 %). La part des hôtes non échantillonnés dans la contribution au risque de la maladie devrait être principalement due aux oiseaux porteurs des *B. garinii*.

ARTICLE 3

High-throughput sequence typing characterization of vector-borne genotypes of *Borrelia burgdorferi* s.l. gives multiple lines of evidence to study transmission cycles and zoonotic risk in a French peri-urban forest

Maude Jacquot, David Abrial, Severine Bord, Maud Marsot, Sébastien Masegla, Angélique Pion, Valérie Poux, Jean-Louis Chapuis, Nathalie Boulanger, Gwenaél Vourc'h, Xavier Bailly

Article en préparation.

High-throughput sequence typing characterization of vector-borne genotypes of *Borrelia burgdorferi* s.l. gives multiple lines of evidence to study transmission cycles and zoonotic risk in a French peri-urban forest

AUTHORS AND AFFILIATIONS

Maude Jacquot¹, David Abrial¹, Patrick Gasqui¹, Severine Bord¹, Maud Marsot^{1,*}, Sébastien Maseglier¹, Angélique Pion¹, Valérie Poux¹, Jean-Louis Chapuis², Nathalie Boulanger³, Gwenaél Vourc'h¹, and Xavier Bailly¹

1: INRA, UR346 Épidémiologie animale, Saint Genès Champanelle, France

2: MNHN-CNRS-P6, Département Ecologie et Gestion de la Biodiversité, UMR 7204 Conservation des Espèces, Restauration et Suivi des Populations, 61 rue Buffon, 75231 Paris Cedex 05, France

3: EA7290: Virulence bactérienne précoce, Groupe Borréliose de Lyme, Centre National de Référence Borrelia, Université de Strasbourg, Strasbourg, France

* present adress: ANSES, Unité Épidémiologie, 23 avenue du Général de Gaulle, 94706 Maisons-Alfort Cedex, France

ABSTRACT

Lyme disease is a zoonosis caused by various species belonging to the *Borrelia burgdorferi* bacterial species complex. These pathogens are transmitted by ticks and infect multiple host species. From an epidemiological point of view, it is important to determine how genetic variants within the species complex are circulating through ticks and in host communities. To this end, we characterized the genotypes of members of the *B. burgdorferi* species complex from 488 infected ticks sampled in a peri-urban forest. For each infected tick, we used a high-throughput approach to sequence the housekeeping gene *rplB*, and the infection-related gene *ospC*. These genotypes were compared to: i) those previously observed in 228 infected rodents belonging to three different species, which were sampled in the same study site; ii) genotypes described in the literature. This allowed us to evaluate the amount of genetic diversity both within and among species in the *B. burgdorferi* species complex in ticks and rodents. Based on the patterns of association of genotypes within infected samples, we identified four main transmission cycles of the bacteria. The spatial distribution of bacterial genotypes assigned to different transmission cycles are not homogeneous, raising question about their ecology and their zoonotic potential. Based on a statistical model, we showed that different transmission cycles, which involve either chipmunks or birds, generate most of the infected nymphs, while the cycle involving bank voles contribute for less than 5% to tick infections. Genotypes assigned to both of the most important transmission cycles were isolated from patient affected by Lyme disease which were infected in this area, suggesting a negative situation from an epidemiological point of view.

INTRODUCTION

A large part of the pathogens that infect humans are zoonotic, *i.e.* are maintained in animal populations [Woolhouse and Gowtage-Sequeria, 2005], and a significant fraction of them are vector-borne pathogens [Jones *et al.*, 2008]. More than 70% of zoonotic infectious diseases originated from wildlife and most of them are due to pathogens that maintain in multiple host species [Cleaveland *et al.*, 2001; Woolhouse *et al.*, 2001]. Indeed, such pathogens can infect multiple host species which altogether contain enough susceptible and competent individuals to ensure their maintenance, and constitute the pathogen's reservoir [Haydon *et al.*, 2002; Viana *et al.*, 2014]. Heterogeneities in the transmission rate of the pathogen genotypes among host species may delineate different transmission cycles in which they can circulate.

As a consequence, describing the genetic diversity of zoonotic pathogens is a key to untangle the circulation of the pathogen. Multilocus sequence typing (MLST) based on Sanger sequencing became a “gold standard” to describe bacterial diversity during the last fifteen years [Maiden *et al.*, 1998]. Nevertheless, the application of this method to the in-vivo characterization of bacterial pathogens that cause mixed infection is limited by technical constraints. New sequencing technologies open opportunities to circumvent these problems, allowing to characterize in detail the genetic diversity of pathogens within infected individuals and to perform studies at the population scale. The genetic information provided by these new methods would contain different lines of information on the way pathogens spread and are maintained through host species and vectors [Viana *et al.*, 2014]. Frequencies of pathogen genotypes among the different host species or vectors vary according to the intensity of pathogen transmission among these epidemiological compartments. Furthermore, barrier to transmission and interactions between pathogens (*e.g.* facilitation), will result in patterns of co-occurrence of bacterial genotypes specific to epidemiological compartments [Vaumourin *et al.*, 2014]. In parallel, the spatial distribution of pathogen genotypes could be used to obtain more information as it is supposed to be influenced by the behavior of infected species. Finally, the genetic information obtained from pathogens infecting ticks and isolated from human cases can be exploited to study the risk for human health due to the different circulating genotypes. In fact, data obtained from questing ticks give a picture on the exposure of human to pathogens and the comparison of tick-borne genotypes and genotypes causing human case can illustrate further their zoonotic potential, especially when no experimental

Chapitre 3

test for virulence in humans is available. In order to reduce the risk of human infection caused by multi-host vector-borne pathogens, describing pathogen diversity in different epidemiological compartments, defining the reservoir(s) of the observed genotypes, and exploring their zoonotic potential are important steps; it can bring keys to develop preventive measures or to manage the host community.

The *B. burgdorferi* species complex, which includes the causing agent of Lyme disease, is an ideal study case to explore these issues. In Europe, these bacteria are transmitted by a generalist tick vector, *Ixodes ricinus*, to various vertebrate hosts. As the culture of these bacteria is fastidious, epidemiological surveillance has been performed by MLST typing on infected ticks without prior bacterial isolation based on Sanger sequencing [Margos *et al.*, 2008]. Yet, different studies reported evidences of co-infection within ticks [Guttman *et al.*, 1996; Qui *et al.*, 2002; Crowder *et al.*, 2010; Tsao *et al.*, 2013]. Even though a HiMLST protocol based on high-throughput sequencing has been developed for these bacteria [Jacquot *et al.*, 2014], it has not been applied to infected ticks yet. In such experimental setting, the precision of sequence data provides an accurate picture of genetic diversity, while the multiplicity of sequences can be used to resolve co-infections. Moreover, data obtained on different epidemiological compartments could be used to study the reservoirs of the different genotypes and their impact on the bacterial population which infects vectors.

The frequency of different alleles at a bacterial infection-related gene, which encodes the outer surface protein C (*ospC*), has been shown to differ among host species [Brisson and Dykhuizen, 2004] and this pattern can extend to markers located in house-keeping genes [Jacquot *et al.*, 2014]. These variations could be used to identify the host range of genotypes belonging to the *B. burgdorferi* species complex. Moreover, studies of co-infection of bacterial genotypes in infected ticks at the genospecies level, based on hybridization protocols, suggest that vertebrate host species structure the distribution of genotypes within ticks [Herrmann *et al.*, 2013]. Thus, the combination of sequence data obtained from hosts and ticks is supposed to give insights into the different reservoirs of *B. burgdorferi* sensu lato (s.l.) genotypes.

Then, estimating the contribution of each reservoir to human population exposure to Lyme disease risk is an important step to plan control measures. Different models were

developed to characterize this contribution (*e.g.* [Brisson and Dykhuizen, 2006; Brisson *et al.*, 2008]). These different models take advantage of differences in either the frequency or the transmission rate of pathogen genotypes in different host species to estimate the proportion of infected vectors generated by each host species. Yet, the parametrization of the most complex of these models used large amount of data: the genotype of pathogens found in questing vectors, the density of the different host species, their specific tick burden, and the transmission probabilities of each bacterial genotype from each host species to vectors. While measuring the contribution of different hosts to the circulation of the *B. burgdorferi* s.l. genotypes, only a few of these genotypes are able to induce Lyme disease in humans [Rudenko *et al.*, 2013; Stanek and Reiter, 2011]. Surveillance surveys of genotypes that are found in both vectors and humans indeed revealed a stronger zoonotic potential of some genotypes [Dykhuizen *et al.*, 2008]. This information should be taken into account to optimize preventive measures.

In this paper, we used HiMLST to explore the genetic diversity of bacteria of the *B. burgdorferi* species complex in questing ticks sampled in a forest located in the suburb of the French capital. Based on these data and former information about the genotypes circulating in different host species, the aims of this paper were i) to assess the applicability of high-throughput sequencing in epidemiological surveillance, ii) to characterize genotype transmission cycles by analyzing their occurrences in hosts and vectors, iii) to evaluate the contribution of different host reservoirs to the distribution and abundance of the different genotypes in ticks using a modeling approach, and iv) to check whether the identified genotypes matched with those obtained from a few reported human cases. Results are discussed in the context of current knowledges of molecular epidemiology of the *B. burgdorferi* species complex and other vector-borne diseases.

MATERIALS AND METHODS

In this study, we sampled ticks, selected individuals infected with *B. burgdorferi* s.l. or related *Borrelia* genotypes, and sequenced two bacterial genes. Genotypes at each locus were identified and compared with those previously identified in infected rodents [Jacquot *et al.*, 2014].

Ticks sampling and *Borrelia* loci sequencing

Ethic statement

Our field study did not involve any endangered or protected species. No specific permission is required by French law to sample ticks and to perform field studies in these locations. Nevertheless, tick sampling has been performed under the supervision of members of the French National Forest Office, who manages the sampling site. Two strains isolated from human cases due to infection events which occurred in the Sénart forest were obtained from the French National Research Center on *B. burgdorferi* located in Strasbourg.

Sample selection

I. ricinus ticks were caught during a field sampling performed in 2011 in the forest of Sénart (3200 ha, 48°40'N, 02°29'E), located near Paris (France), by dragging a large piece of cotton fabric, called flag hereafter, across the vegetation and leaf litter.

The Sénart forest is divided in 333 plots, which were surveyed for ticks. Wherever possible, the flag has been dragged over two lines of 80 meters in each plot, checking for the presence of ticks every 10 meters. At least, one nymph tick has been obtained from 413 lines. For each tick developmental stage (*i.e.* larvae, nymphs, and adults), the sampled ticks from each line were pooled in a tube containing 70% ethanol, and stored until further analyses. Among these ticks, 3903 were selected at random, resulting in a sampling scheme in which, at least, one tick was included for each positive flag dragging line.

For each tick, DNA has been extracted using the ammonia based protocol described by Humair *et al.* [2007]. We then looked for *Borrelia* DNA in these extracts using a quantitative PCR protocol. With this aim, amplifications of a part of the *flaB* gene were performed using two primers : FlaB_outF 5'-ATATAACCAAATGCACATGTT-3', and FlaB_inR 5'-ACATTAGCWGMATAAATATTTACAG-3'. The amplification has been performed with the SYBR Green SsoAdvanced Supermix kit on a CFX Touch Detection System (Biorad, Hercules, California, USA). Amplification was performed in a total reaction volume of 20µl

including 10µl of 2X mix, 2µl of each primer at 10µM, 1µl of H₂O, and 5µl of DNA template solution. The amplification started with an initial denaturation step of 2 min at 98°C, and 50 cycles including a denaturation step of 5s at 98°C and an annealing/elongation step of 30s at 60°C, recording fluorescence at each cycle.

Sample characterization

As the aim of this study was to evaluate, in depth, the genetic diversity present within infected ticks, we selected, among all individuals infected by *B. burgdorferi* s.l. 488 ticks at random. DNA samples from isolates of the *B. burgdorferi* species complex, cultured from two humans infected in the Sénart forest, *i.e.* IBS18 and IBS42, were added to the experiment. We studied the genetic diversity of *Borrelia* genotypes at two different loci, as performed previously [Jacquot *et al.*, 2014]. In summary, we focused on partial sequences of the housekeeping gene *rplB* and the infection-related gene *ospC*. Markers were amplified independently *via* semi-nested PCR for each DNA sample using the Promega GoTaq kit (Madisson, Wisconsin, USA), using sample specific tagged primers (Sigma-Aldrich, Saint-Louis, Missouri, USA). Equal amounts of purified *rplB* and *ospC* PCR products were then mixed with each other and sent to GATC-Biotech (Konstanz, Germany) for sequencing on three eighths of a GS FLX+ run.

Sequence analysis

Rodent sequences

One aim of the study was to characterize the transmission cycles of *B. burgdorferi* s.l. genotypes through vectors and host. Thus, for sequences analysis, we decided to use *rplB* and *ospC* sequences we obtained from ticks for this study and those we previously obtained from 228 infected rodent [Jacquot *et al.*, 2014]. The rodent sample included 125 chipmunks (*Tamias sibiricus barberi*), 93 bank voles (*Myodes glareolus*), and 10 wood mice (*Apodemus sylvaticus*). All further analyses included sequences obtained from both ticks and rodents.

Assignment of sequences to individuals and loci, trimming and alignments

Based on the method that was described in Jacquot *et al.* [2014], raw sequences were assigned to their respective tick or host individuals and targeted loci using a BLASTN approach [Altschul *et al.*, 1990]. The tag and the targeted sequence from each raw read were compared to a database that contained respectively a list of the oligonucleotides used to tag

Chapitre 3

sequences obtained from each sample, and different representative sequences of the targeted loci extracted from the *B. burgdorferi* MLST database (<http://borrelia.mlst.net>) [Margos *et al.*, 2008] and GenBank [Benson *et al.*, 2013]. We used only the first 500 bp of sequences for subsequent analyses and sequences shorter than 350 bp were removed from the dataset. For each individual, we constructed an alignment of the remaining sequences for each locus using K-Align [Lassman and Sonnhammer, 2005].

Genotype delineation and phylogeny

The main disadvantage of high-throughput sequencing methods, compared to Sanger sequencing, is the high base-calling error rate. Studying diversity, this could be problematic due to the introduction of low-frequency variants that are artifacts of sequencing error. In order to circumvent this problem, genotypes were delineated according to two successive distance-based nearest-neighbor classifications: i) at the individual level and ii) for the whole dataset [Jacquot *et al.*, 2014]. According to the diversity pattern of each marker and for the both classifications, thresholds for divergence were fixed. We chose thresholds of three and four nucleotide sites for *rplB* and *ospC* respectively in order to: i) limit the number of groups delineated by the algorithm; ii) keep a resolution level that do not affect results of further analyses. Genotypes containing fewer than four sequences were excluded from subsequent analyses.

We used a phylogenetic approach to examine the diversity of the observed genotypes in further details. For each locus, consensus sequences and reference sequences, were aligned with each other using K-Align [Lassman and Sonnhammer, 2005]. Phylogenetic searches were performed with a maximum likelihood approach using PHYML [Guindon and Gascuel, 2003]. The most appropriate model of evolution was chosen for each alignment according to the Akaike Information Criterion [Akaike, 1974] which was calculated using the APE library in R [Paradis *et al.*, 2004]. Then, distance matrices were computed from alignments with PAUP* 4.0 b10 [Swofford, 2003] using the best model for each locus, *i.e.* TN93+I+G for *rplB* and GTR+I+G for *ospC*. The substitution rate matrices were estimated *via* maximum likelihood and we assumed empirical nucleotide frequencies. As calculated by the PHYML analysis, the shape parameters of the gamma distributions were fixed at 0.663 for *rplB* and 0.958 for *ospC*, the proportion of invariable sites were fixed at 0.410 and 0.233, for *rplB* and *ospC* respectively. At that point, for both loci, phylogenetic networks were obtained with the Neighbor-Net method [Bryant and Moulton, 2004] using SPLITSTREE4 [Huson and Bryant,

2006].

Delineation and co-occurrence of genotype groups within ticks and host species

Using the networks, we were able to empirically delineate groups of closely related genotypes; *i.e.* genotype groups (GG). GG were numbered according to the previously established nomenclature [Jacquot *et al.*, 2014].

First, we aimed at identifying the different transmission cycles of the *B. burgdorferi* s.l. GG we found in ticks and rodents. We built a co-occurrence matrix that described the amount of co-occurrence of GG within individuals. To obtain this information we used an incidence normalized matrix ($m \times n$), where m was the number of tick and host individuals and n was the number of GG. After, we described the presence or absence of the different GG among sample individuals and used the normalized incidence matrix ($m \times n$) to create a co-occurrence matrix ($m \times m$) that described the GG that each individual shared with the others. This matrix was obtained by multiplying the incidence matrix by its transpose. Next, we evaluated which tick and host individuals showed similar patterns of infection by *B. burgdorferi* s.l. GG, *i.e.* sets of particularly connected individuals called communities in a statistical context, using a “greedy” approach [Clauset *et al.*, 2004]. To identify these infection groups (IG), we chose the classification that optimize the modularity of the graph by: i) maximizing a modularity index, and ii) classifying samples among the lowest number of communities. From this, we calculated the proportion of individuals among ticks and each host species that were classified in the different communities. Finally, we summarized this information by building a meta-graph that: i) discriminates the different communities, ii) distinguish ticks and rodents within communities.

Similarly and with the aim to put forward multi-locus associations, *i.e.* GG at both markers that appeared in the same hosts, we created a co-occurrence matrix ($n \times n$) describing the number of hosts and vectors in which each GG was shared. This matrix was obtained by multiplying the incidence matrix and its transpose. The clustering algorithm described above [Clauset *et al.*, 2004] allowed us to illustrate the associations among GG, *i.e.* sets of genotype groups (SGG) infecting the same tick and host individuals, with a dendrogram.

These analyses were performed using the igraph library in R [Csardi and Nepusz, 2006]. The similarity matrix was obtained from the incidence matrix using the graph.incidence and the bipartite.projection functions. The classification analyses were performed using the fastgreedy.community function, and graphical outputs were produced

Chapitre 3

using the `tkplot` and `dendPlot` functions.

Infection groups mapping and spatial distribution

To visualize the spatial distribution of the IG in the forest, we built a map using QGIS 2.4. As ticks sampled in the same line were pooled, the exact localization of each tick was not known so we had to aggregate data to the centroid of the sampling lines. Thus, on one location it was possible to observe more than one IG. To display clearly this diversity we plotted on the centroids a pie chart which represented the ratio of each IG present on the sampling line.

In the following step, we performed a test to investigate if the different IG were randomly and uniformly distributed over the forest of Sénart. To this end, we estimated a $\hat{G}(h)$ function independently for each IG, which describes the cumulative distribution of distances (by steps of 0.5 meter) between tick individuals belonging to a given IG and the closest tick belonging to the same IG [Bivand *et al.*, 2013]. Estimates of the functions, were compared to an envelope obtained by the following Monte-Carlo approach. One thousand simulated datasets were obtained, according to the sample size of the studied IG, sampling ticks randomly from all available infected ticks. Envelopes were based on the 2.5 and 97.5% quantiles of empirical $\hat{G}(h)$ functions calculated from the simulated datasets. To assess the impact of sampling constraints and biological characteristics on the observed spatial distribution, we also obtained a theoretical distribution of $\hat{G}(h)$ assuming a Poisson spatial point process of intensity λ , corresponding to the density of ticks belonging to the studied IG.

Contribution model: host contributions to ticks diversity

We developed a statistical model to estimate the relative contribution (in percentage), to the number of nymphs infected by *B. burgdorferi* s.l., of chipmunk-associated bacterial GG, bank-vole-associated GG, and GG not associated to the studied hosts. Additionally, the model can be used to determine if additional, non-sampled, vertebrate species should be considered as important hosts for *B. burgdorferi* s.l.. Such model could also be used to predict which vertebrate species is the host of human infectious strain among the whole host community of *B. burgdorferi* s.l..

Our model derives from the “signature matching” and the “inverse model” developed by Brisson *et al.* [2008] to deal with the specificity of our dataset. For instance, we did not have access to transmission probabilities, from each host species to ticks, of the different

B. burgdorferi s.l. GG; probabilities which are required for the “inverse model”. Because of this lack of information, we based our analysis on the similarity of bacterial genotypes in infected hosts and ticks, as in the “signature matching” model. To avoid the use of rule-based assignation of ticks to infer contribution, as performed in the “signature matching” model, we used a simulation approach similar to the method used for the “inverse model”.

By definition, the infection patterns of unstudied hosts, called the X category hereafter, were unknown. To avoid situations where the inferred infection patterns of the X category match completely the infection patterns observed in ticks, and is considered as the only significant contributor, we simulated tick infection pattern in a sequential way.

The first sequence aimed at defining a distribution of the studied hosts contributions, without taking into account the X category. Thus, these distribution corresponds to the maximal contribution of these hosts to the modeled pathosystem. For each simulation, three values α , β and γ , corresponding to the respective contributions of chipmunks, bank voles and the X category of host species to infected ticks, were obtained. These contributions were generated randomly so that $\alpha + \beta + \gamma = 1$.

Based on observed data, we scored individuals according to the presence of alleles belonging to the different SGG. We then used the observed frequencies of combinations of SGG presence/absence patterns, i , in chipmunks and bank voles, C_{iT_s} and C_{iM_g} respectively, to compute frequencies of combinations of SGG presence/absence patterns in ticks so that $C_{iT_kSim} = \alpha C_{iT_s} + \beta C_{iM_g}$. Using combinations of SGG presence/absence patterns allow us to explicitly take into account i) the information brought by both markers and ii) co-infection patterns. One million C_{iT_kSim} were simulated and compared to the observed C_{iT_kObs} using euclidean distances between the vectors C_{iT_kSim} and C_{iT_kObs} . The 10% $\{\alpha, \beta, \gamma\}$ values which resulted in the closest simulations to observed data were conserved for the next step.

Starting from selected sets of $\{\alpha, \beta, \gamma\}$ values, combination of SGG presence/absence patterns in ticks were simulated so that, $C_{iT_kSim} = \alpha C_{iT_s} + \beta C_{iM_g} + \gamma C_{iX}$. With this aim, C_{iX} vectors were obtained by: i) drawing a frequency of each SGG from an uniform distribution between 0 and 1; ii) computing the frequency of each combination of SGG presence/absence patterns in ticks based on the frequency of each SGG, assuming the hypothesis of random association of SGG within ticks, and conditioning on ticks being infected by at least one SGG. The sum of all the combination frequencies is thus equal to 1. As in the previous steps, the 100 000 simulations were compared to observations. To this purpose, euclidean distances were build from both: i) C_{iT_kSim} and C_{iT_kObs} vectors, and ii) vectors of differentiation values

Chapitre 3

(Wright's F_{ST} [Wright, 1965]) computed for each SGG between on the one hand ticks and chipmunks and on the other hand ticks and bank voles, in simulated and observed datasets. Sets of parameters were accepted and stored if the distance was lower than a tolerance threshold, defined so that 1% of simulations were accepted.

RESULTS

Raw data

In total, from tick individuals, we obtained 288,151 sequences that were assigned to individual ticks and deposited to the Sequence Read Archive (SRA) database under the accession SRP041191. 75,973 sequences were assigned to a tick individual and to the *rplB* gene and 144,604 sequences were assigned to a tick individual and to the *ospC* gene. Among the 488 nymphs, we obtained, on average, 155.7 *rplB* sequences and 296.3 *ospC* sequences per tick individuals. Even though a few individuals did not yield sequence data, we obtained sequences belonging to 472 different nymphs.

For sequences from rodents, we used the 16,222 sequences that have been deposited in the SRA database under the accession SRP032755 and which corresponds, on average, to 33.2 *rplB* sequences and 37.9 *ospC* sequences per host individual [Jacquot *et al.*, 2014].

Genotype groups and phylogenies

Using a nearest-neighbor classification algorithm, 39 and 82 genotypes were delineated from the sequences for *rplB* and *ospC* loci respectively. Based on phylogenetic networks, we were able to empirically delineate groups of closely related genotypes; *i.e.* genotype groups (GG). Seventeen and 28 GG were delineated for the *rplB* and *ospC* markers respectively (Figures 1 and 2). We calculated the proportion of individuals within ticks and each host species that was infected by any given group of closely related genotypes (Table S1). Tick and host individuals were infected by several *rplB* genotypes, indicating potential co-infection by different species. In fact, 22.6%, 17.8% and 6.5% of tick, chipmunk and bank vole individuals respectively were co-infected by different *rplB* genotypes while 42.3%, 35.6% and 35.7% of ticks, chipmunks and bank voles individuals respectively were co-infected by multiple *ospC* genotypes (Table 1).

Table 1. Percentage (%) of ticks and host individuals of chipmunks and bank vole in which co-infection by at least two genotypes of *Borrelia burgdorferi* s.l. and/or *B. miyamotoi* detected.

Host species	<i>rplB</i>	<i>ospC</i>
<i>Ixodes ricinus</i> (ticks)	22.60	42.81
<i>Tamias sibiricus barberi</i> (chipmunks)	17.78	35.56
<i>Myodes glareolus</i> (bank voles)	6.52	35.72

Using reference sequences, the *rplB* neighbor-net network showed that many species

Chapitre 3

of the *B. burgdorferi* species complex infected ticks and rodents in the Sénart forest (Figure 1): *B. burgdorferi* s.s. with the *rplB* G1 GG; *B. spielmanii* with *rplB* G2; *B. afzelii* with *rplB* G3 and G4; *B. garinii* with *rplB* G5 and G8; *B. valaisiana* with *rplB* G7 and G10; *B. lusitaniae* with *rplB* G13. New genotypes, which did not cluster with any known sequences, were also found. More specifically, the GG *rplB* G6, G11, G12 and G14 could hardly be assigned to a species. Even if all GG found in rodents were also represented in ticks, few GG were mainly carried by nymphs. The *rplB* G1 and G4 GG, which have been shown to be associated with chipmunks [Jacquot *et al.*, 2014], infected 93 and 121 nymphs respectively. The *rplB* G3 GG which has been shown to be associated with bank voles infected only 13 nymphs. Moreover, ticks carried more genetic diversity than rodents, as *rplB* G7 and G8, that were almost never found to infect the three rodent species sampled, infected 49 and 27 nymphs respectively. Otherwise, *rplB* G16 and G17 GG that were found in 49 nymphs grouped closely with sequences related to the relapsing fever spirochetes and were particularly close to the two strains of *B. miyamotoi* FR64b and LB-2001.

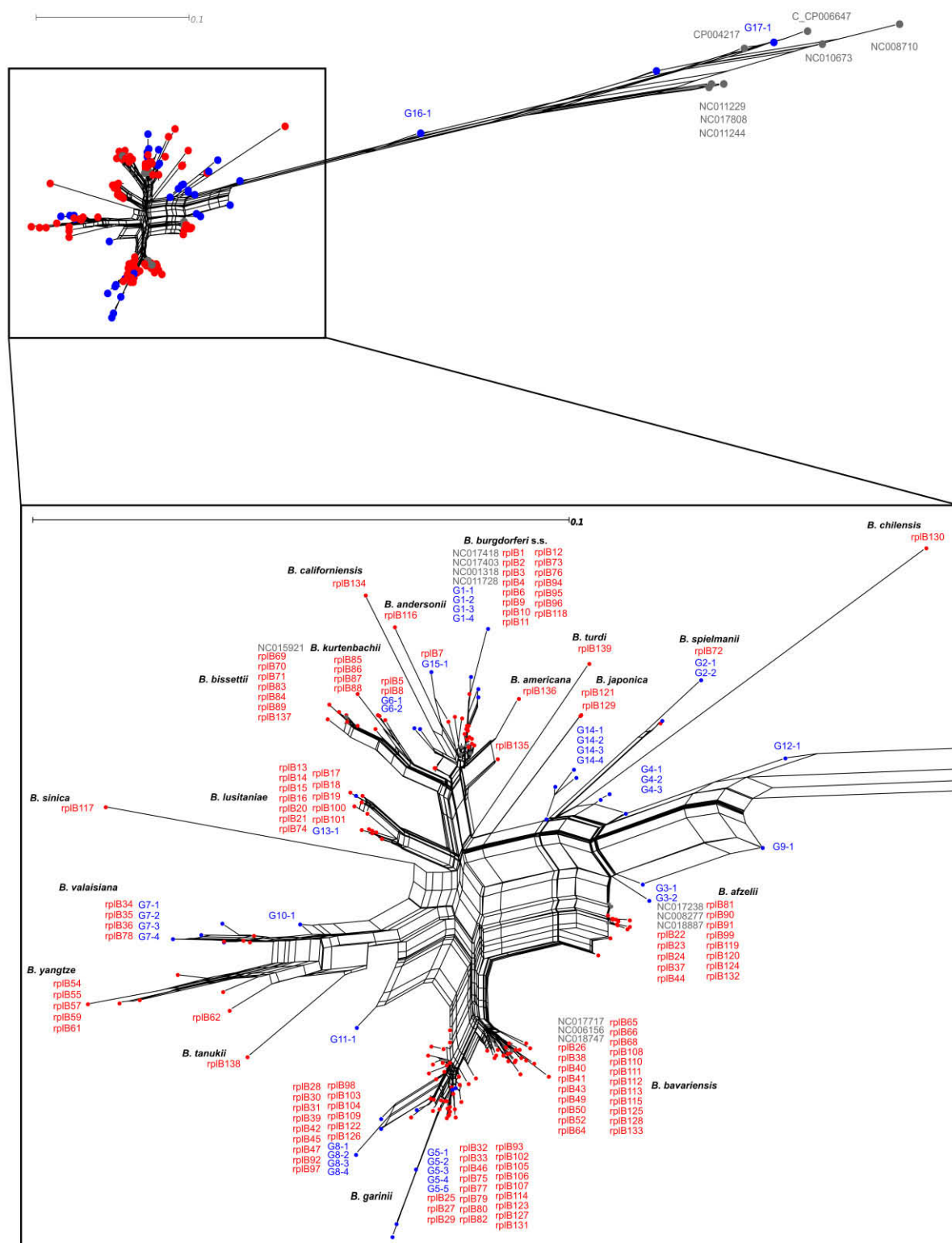


Figure 1. Phylogenetic networks of *rplB* sequences.

The network includes consensus sequences of the genotypes recovered in this study (in blue) as well as reference sequences for species of the *B. burgdorferi* species complex and from the relapsing fever (in red and gray). All genotypes having the same prefix (e.g. G6-) were empirically assigned to the same genotype group because of their close relationships.

Chapitre 3

The phylogenetic network of *ospC* genotypes revealed a higher genetic diversity - compared to what can be observed at the *rplB* locus- and presented a star-like layout (Figure 2). Among the 27 genotypes listed for the *B. burgdorferi* s.s. species and annotated by capital letters [Rudenko *et al.*, 2013], 7 were represented in our sampled. Twelve GG identified in this study were found only in infected nymphs (Table S1). Among *ospC* GG, the most frequently found in ticks were those carried by chipmunks as for example *ospC* G1, G10, G11, G14. Conversely, *ospC* G6, G7, G21 GG were each found in more than 35 nymphs but only in few or none rodent individuals. Additionally, the *ospC* G3 and G8 GG were respectively represented in 55 and 37 bank voles that were sampled but, seemed to infect only a few nymphs (12 and 36, respectively).

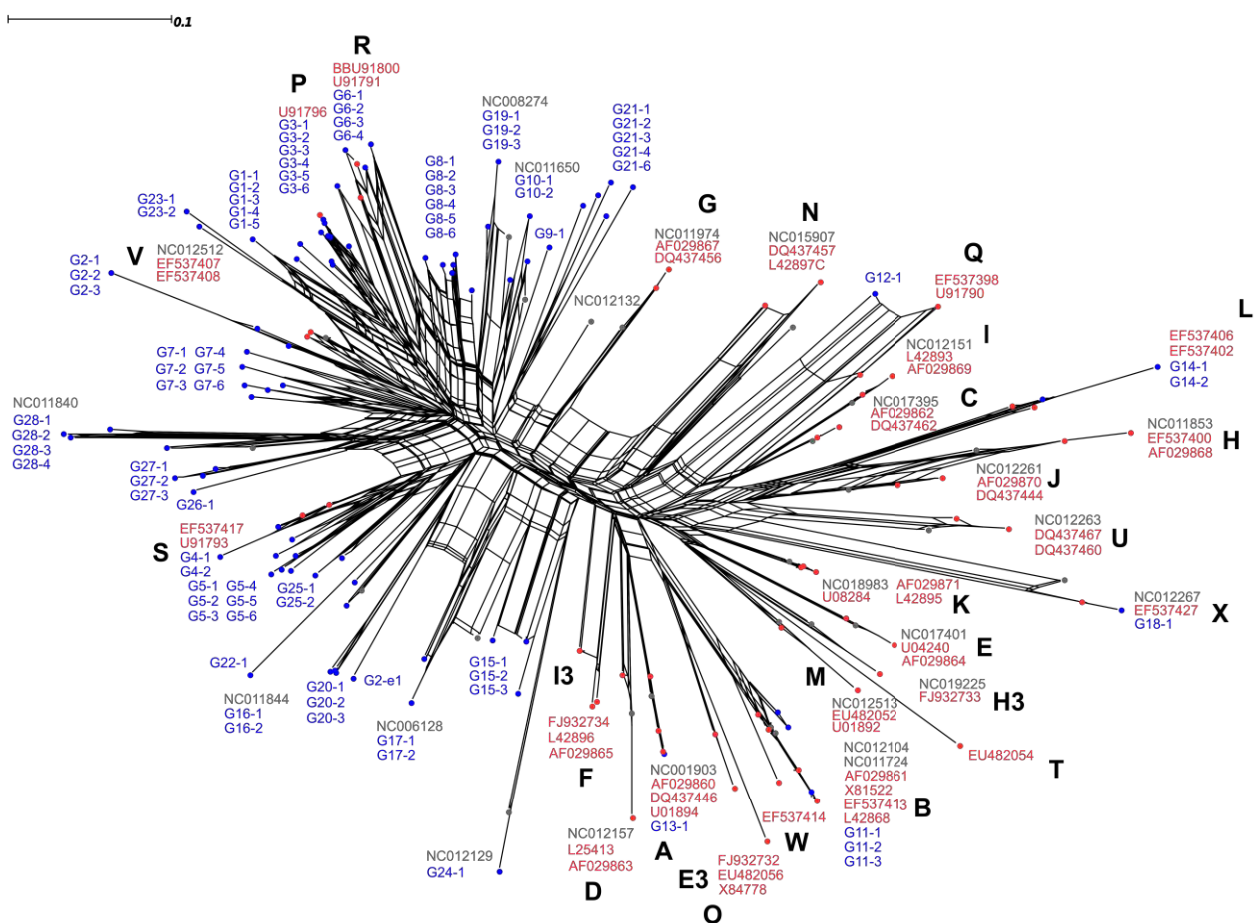


Figure 2. Phylogenetic networks of *ospC* sequences.

The network includes consensus sequences of the genotypes recovered in this study (blue) as well as reference sequences for species in the *Borrelia burgdorferi* species complex (red and gray). Capital letters to the side of groups of genotypes represent unique *ospC* alleles described in previous studies [Rudenko *et al.*, 2013]. All genotypes having the same prefix (e.g. G21-) were empirically assigned to the same genotype group because of their close relationships.

Among the two human strains we characterized, the IBS18 strain was assigned to *B. garinii* based on the *rplB* sequences we obtained. These sequence were similar to the *rplB* G5 GG. The *rplB* sequences obtained from IBS42 suggested that this strain belong to *B. afzelii*. Its sequences were related to the *rplB* G4 GG (data not shown). Their *ospC* sequences were respectively close to *ospC* G16 and *ospC* G10 GG.

Individuals sharing similar genotype groups infection pattern

We classified tick and host individuals we sampled based on the bacterial GG they shared using a graph approach. This analysis showed that 7 sets of individuals sharing the same kind of diversity patterns, *i.e.* infection groups (IG), can be distinguished from each other. The obtained results were summarized in a meta-graph (Figure 3). Two IG mainly included chipmunks (50 and 44) and ticks (115 and 107) and a few other rodents (1 bank vole and 4 wood mice in the last). This correspond to ticks and chipmunks which were mainly infected by the *rplB* G1 and G4 GG respectively; GG appearing to belong to *B. burgdorferi* s.s. and *B. afzelii* (Figure 1). Results obtained from sets of genotype groups (SGG) research showed that many GG of *ospC* also infected these individuals (Figure 4). A third IG was composed of 89 bank voles, 4 wood mice and only 18 ticks. Two others IG, highly represented in individuals we sampled, were mostly composed of ticks (102 and 37). The first of them was predominantly composed of individual carrying *rplB* GG -G5, G7, G8 and G10- which were closely related to *B. garinii* and *B. valaisiana* sequences (Figure 1). The other one was composed of nymphs infected by the *rplB* G17 GG which was close to *B. miyamotoi* sequences. The main associated *ospC* GG are shown Figure 4. The two last IG were disconnected from the rest of the meta-graph and only composed of 2 and 1 nymph(s).

Infection groups spatial distribution

To get further insights into the spatial distribution of individual infection patterns within Sénart, we mapped the different IG ticks belonged to in function of their sampling location (Figure 5). Most of IG were homogeneously distributed in the Sénart forest, excepted the IG1 and IG4 which were significantly over-aggregated in perimeters of less than 400 meters (Figure 6). This means that ticks belonging to these IG were separated by distance lower than it would be expected based on the spatial distribution of all infected ticks.

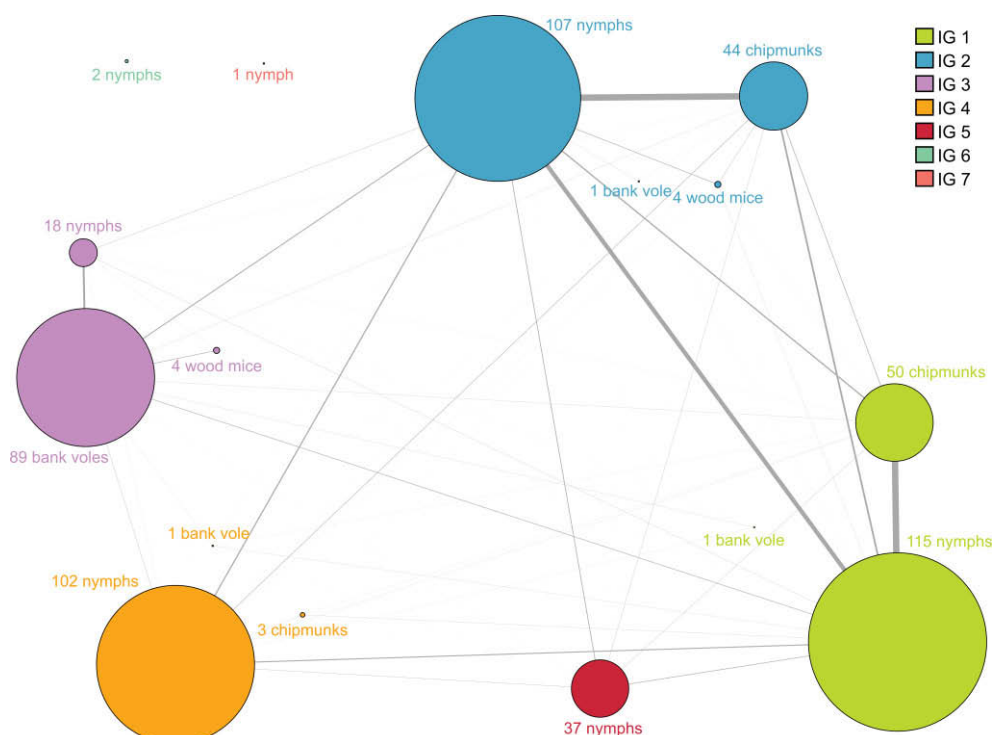


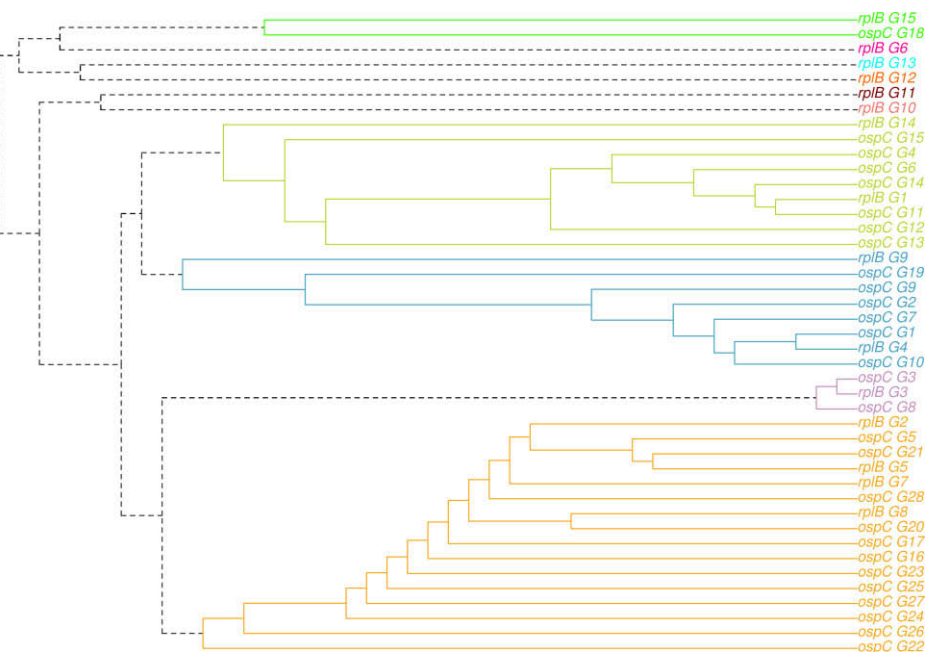
Figure 3. Graphical representation of infection groups compositions

The graph includes all tick and host individuals for which sequences from at least one of the *rplB* or *ospC* genotype group was obtained. Each distinct community of individuals that were infected by a similar infection pattern (*i.e.* IG) that was identified, using a “greedy approach”, is highlighted in a different color and can be composed of one or more host and tick individuals. Circle were placed empirically and link widths are proportional

to the number of genotypes shared by the individuals of pairs of meta-groups.

Figure 4. Patterns of genotype groups associations.

The



dendrogram includes genotype groups of both targeted loci, *rplB* and *ospC*, detected in ticks and host individuals excluding the *B. miyamotoi* *rplB* groups G16 and G17. Communities of frequently co-occurring genotype groups, *i.e.* SGG, were identified using a greedy approach and are here highlighted by different colors.

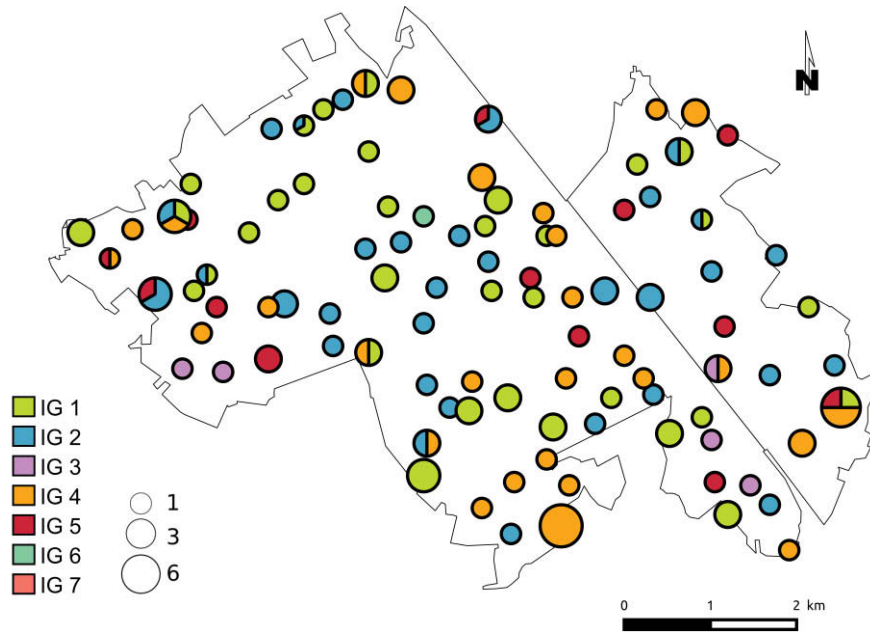


Figure 5. Infection groups spatial distribution

On this Sénart forest map, the spatial distribution of the infection groups (IG) is illustrated. Communities of individuals that are infected by a similar infection pattern, *i.e.* IG, were identified using a “greedy approach”. The seven IG identified are represented by different colors. Pie charts show the ratio of each IG present on each sampling line. Diagram sizes are proportional to the number of infected nymphs.

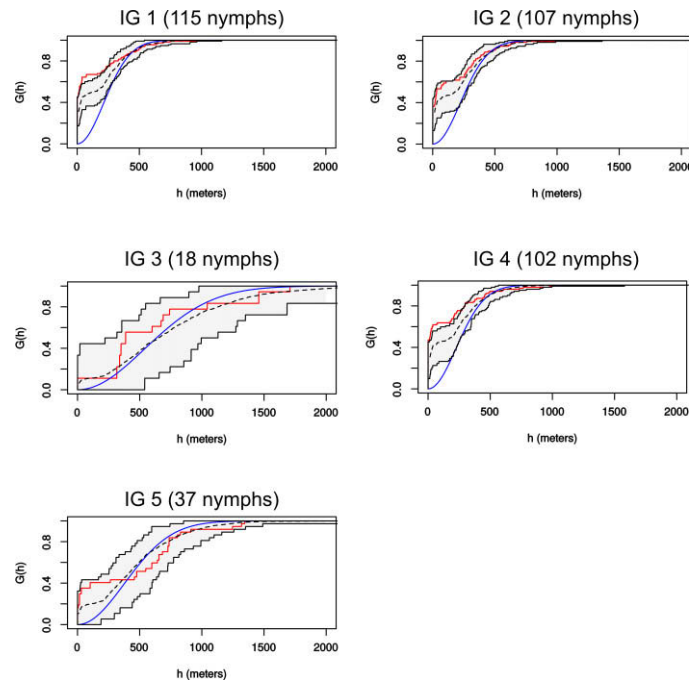


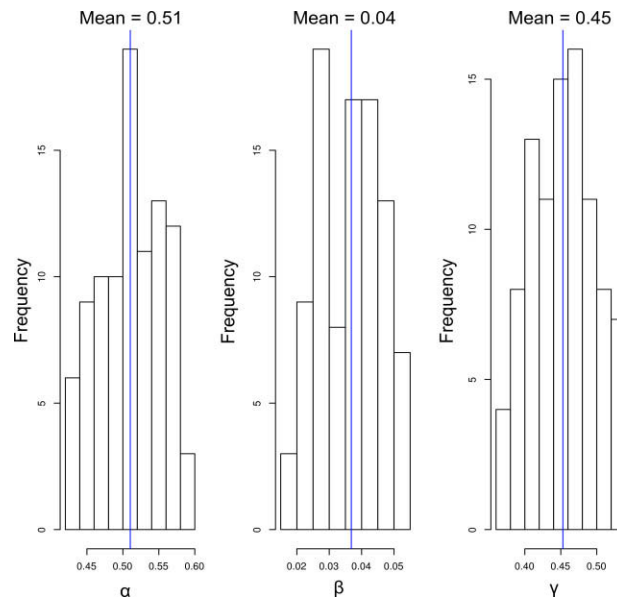
Figure 6. Infection groups events distance distribution

The red lines describe empirical $\hat{G}(h)$ functions, *i.e.* cumulative distributions of the shortest distance between each occurrence of IG and similar infection events, per steps of 0.5m. The 95% envelopes of confidence were built from 1000 simulated $\hat{G}(h)$ functions based on random samples of infected ticks. The black dashed line is the mean of the envelope for each values of the distance h . In blue, theoretical $\hat{G}(h)$ distributions were obtained using a Poisson random point process.

Chapitre 3

Host contributions to infected ticks

As we aimed at estimating relative contributions of chipmunks, bank voles and non sampled hosts to the number of nymphs infected by *B. burgdorferi* s.l. genotypes, we identified SGG in our dataset excluding the two *B. miyamotoi rplB* G16 and G17 GG. GG at both genes did not infect tick and host individuals randomly as 10 SGG have been found when analyzing their co-occurrence patterns (Figure 4). First, the *B. afzelii rplB* G3 GG is associated with both *ospC* G3 and G8, mainly found in ticks and/or in bank voles. The *B. spielmanii*, *B. garinii* and *B. valaisiana rplB* G2, G5, G7 and G8 GG seemed to infect the same individuals and were associated with numerous *ospC* GG that were only found in ticks (Table S1). The third and fourth SGG grouped together the *B. afzelii rplB* G4, G9 and *rplB* G1 and G14 GG (*B. burgdorferi* s.s. and non assigned species) with *ospC* GG that infected mainly ticks and/or chipmunks. The *rplB* G15 grouped with an unique *ospC* GG, *ospC* G18, whereas the *rplB* G6, G10 G11 G12, G13 groups, did not cluster with any other GG defined in the dataset. To estimate the relative host contributions to the number of tick, we simulated data that we compared to the empirical frequencies of these 10 SGG combinations. From posterior distributions of the parameters of interest (Figure 7), we computed average values for each parameter. The resulting estimates of the contribution parameters were $\alpha = 51.0 \%$, $\beta = 3.7 \%$ and $\gamma = 45.3 \%$ which correspond to the contribution of the host reservoirs



respectively including chipmunks, bank voles and the X category.

Figure 7. Distribution of the parameter values of the contribution model.

Distribution of the parameters of interest values for the simulations we selected (based on their similarity with our observations) are plotted. α , β and γ correspond to chipmunks, bank voles and X category contributions'

respectively.

DISCUSSION

Here, we applied high throughput sequencing to two markers of the *Borrelia burgdorferi* species complex to efficiently characterize the genetic diversity of the complex within tick vectors and the spatial distribution of this diversity at a forest scale. This study particularly shows the importance of taking account co-infections patterns to better understand the genetic structures that can be observed and the way pathogen spread and are maintained. In a second time, we compared this diversity to genotypes at both genes previously described in three rodents species captured in the same forest, from 1 to 6 years ago and we evaluated the contributions of sampled and non-sampled host species to ticks infected by *B. burgdorferi* s.l..

Many species of the complex were present in our tick sample. Many nymphs were infected by *B. burgdorferi* s.s., *B. afzelii*, *B. garinii* and/or *B. valaisiana*. Less frequently, we also found genotypes close to those of *B. spielmanii* or *B. andersonii*. Genospecies co-infections in ticks were frequent, a result which is congruent to what have been previously observed in Europe for *I. ricinus* [Rauter and Hartung, 2005; Herrmann *et al.*, 2013]. We also found sequences (*rplB* G16 and G17) close to *B. miyamotoi* sequences, a member of the *Borrelia* relapsing fever species complex, suggesting that a typing scheme targeting both of these sympatric emergent pathogen groups might be developed.

Our study demonstrates that, at the Sénart forest scale, the diversity of strains of *B. burgdorferi* s.l. carried by rodents, at both loci *rplB* and *ospC*, is lower than the diversity found in ticks. The higher number of genotypes identified using both tick and host sequences compared with the number of genotypes we previously had identified in rodents can be explained by two non-exclusive assumptions. First, many genotypes we identified in ticks must spread and be maintained in hosts we did not sampled. This is particularly likely for genotypes that explain infection groups (IG) that include only ticks, as the IG4. This IG is characterized by *B. garinii* and *B. valaisiana* genotype groups (GG), *rplB* G5, G7 and G8, which are supposed to be maintained and spread through birds, these species being usually associated with avian reservoirs in Europe (reviewed in Margos *et al.* [2011]). Second, within IG that include rodents, the diversity we observed in ticks might also be influenced either by strains that could be rare in rodents but transmitted during tick co-feeding, or by a higher sequencing effort; on average, we obtained ~5-fold and ~8-fold more sequences for ticks than

Chapitre 3

for rodents individuals at the *rplB* and *ospC* genes respectively.

Considering the result we obtained on the description of both IG, *i.e.* groups of hosts or vectors showing similar infection patterns, and SGG, sets of genotypes at different loci that co-occurred frequently, HiMLST is a promising method to perform the epidemiological survey of *Borrelia* genotypes directly from infected material. In the near future, the resolution of this approach might be improved by increasing the number of loci surveyed to study the spread of pathogens at the strain level [Maiden *et al.*, 2013].

The identification of IG suggests that at least four important different transmission cycles are involved in the maintenance of *B. burgdorferi* s.l. in the Sénart forest. These transmission cycles involves various couple of genospecies of *B. burgdorferi* s.l. and host species: *B. burgdorferi* s.s. and chipmunks, *B. afzelii* and chipmunks, *B. afzelii* and bank voles or *B. garinii* and *B. valaisiana* which are probably associated with birds. These results were to put in front of host contributions that we estimated for sampled and non-sampled host species. It is noteworthy that a supplementary transmission cycle, not included in this quantitative analysis, must be considered for the GG of *B. miyamotoi*.

The statistical model we used allow us to evaluate the relative contribution of sampled and non-sampled host species to infected ticks. It is only based on genotypic data, as we did not have access to ecological measurements as: i) the abundance of host species, ii) susceptibility to pathogens, or iii) transmission rate of *B. burgdorferi* s.l. genotypes from host to ticks and vice-versa. Due to the lack of ecological inputs in this model -see Brisson *et al.* [2008] for a more complex implementation-, it can not give information on critical biological parameters that explain the success of bacterial genotypes in host species. Consequently, the contributions we measured were complex functions of these different parameters. Nevertheless, in an epidemiological context where quantitative contributions of the different host species are of main interest, this approach would be relevant to highlight sources of zoonotic risk.

As in previous modeling attempts, unstudied hosts seem to play an important role in the Sénart forest, *i.e.* more than 40%, in the infection of ticks by the *B. burgdorferi* species complex [Brisson and Dykhuizen, 2006; Brisson *et al.*, 2008]. This illustrates the complexity

of the enzootic cycle of Lyme disease. While figuring out which host species play a important role among unstudied hosts was difficult in the US, because a single bacterial genospecies is present, it is likely in our study that a large part of *B. burgdorferi* s.l. genotypes circulating in unstudied hosts are maintained in birds, based on the reports of *B. garinii* and *B. valaisiana* in these animals [Olsen *et al.*, 1995 ; Humair *et al.*, 1998 ; Kurtenbach *et al.* 1998 ; Hanincova *et al.*, 2003; Dubska *et al.*, 2009]. Indeed, four bird species were identified to contribute to the number of infected ticks in the Sénart forest, the common blackbird, the European robin, the song thrush and the winter wren, but the infecting genotypes were not characterized [Marsot *et al.*, 2012]. Based on the distribution of *B. garinii* and *B. valaisiana* genotypes, a single SGG and a single IG were described, suggesting an unstructured transmission cycle. The co-occurrence of GG within infected ticks and hosts contains for sure underused information on transmission cycles, nevertheless the resolution of the graph approach we used is still unexplored in different epidemiological context, *e.g.* differences in the diversity of pathogens, their prevalence or their transmission mode.

The results obtained for *B. burgdorferi* s.l. GG found in chipmunks illustrate this uncertainty. Two sets of individuals, IG1 and IG2, harbored these GG, which clustered in two corresponding SGG. Co-infections occurred regularly but not at a sufficient rate to create unique and homogeneous sets of both samples and genotypes. Our model suggests that these GG are quite frequent among ticks, with an inferred contribution of chipmunks to infected ticks over 50%. It is coherent with a previous study that demonstrated the strong “reservoir potential” of chipmunks in the Sénart forest [Marsot *et al.*, 2013], even though the number of ticks carrying chipmunks-associated GG seems high regarding the 35% prevalence of the pathogen in chipmunks. The average nymphs burden of chipmunks in this forest exceed 20 [Pisanu *et al.*, 2010], suggesting that an individual could indeed meet around 50 nymphs infected by chipmunk-associated genotypes during a two years long life. This could lead to high prevalence in a Susceptible-Infected model with reasonable transmission probabilities. Explaining the difference between the observed and the expected prevalence in chipmunks could involve two sets of hypotheses: i) low infection rates or strong recovery rates; ii) a lack of resolution in either typing or *in silico* analyses, resulting in the merging of different transmission cycles.

Different hypotheses might explain the delineation of two IG and SGG for chipmunk-

Chapitre 3

associated genotypes. For example, negative interactions among genotypes within hosts, caused by competition or cross-immunity, might affect their distribution by limiting co-occurrence. The two SGG might also maintain in reservoirs including different host species, chipmunks participating to both transmission cycles. This last hypothesis is coherent with the difference in the spatial distribution of ticks that carry the two IG of chipmunk-associated GG. Identifying the role of other host species in the circulation of these GG is important in the perspective of control measures targeting the host community. Experimental evidences obtained from field interventions as vaccination of chipmunks could help to investigate this question.

Conversely, the contribution of bank-voles to infected nymphs was ten times lower than the contribution of chipmunks (~4%). Indeed, the tick burden of bank voles is ~25 times lower than the tick burden of chipmunks [Pisanu *et al.*, 2010]. Despite the low frequency of bank-vole associated genotypes in ticks, the prevalence in this host species reach 16% in the Sénart forest [Marsot *et al.*, 2011], which seems congruent with the probability of bank voles to meet ticks infected by their associated genotypes based on its average tick burden. The low contribution of bank-vole associated SSG to infected ticks suggests a low impact on zoonotic risk. Further than the frequency of genotypes observed in questing ticks, the two isolates obtained from humans infected in the Sénart forest were related to the genotypes associated with i) unstudied hosts, probably birds, and ii) chipmunks, illustrating a negative epidemiological situation. Moreover, the *ospC* G13 and G11 GG, carried by respectively 5 and 22 chipmunks and 0 and 43 nymphs are closely related to the *ospC* major group A and B defined by Seinost and colleagues [Seinost *et al.*, 1999] as alleles of *B. burgdorferi* s.s. that causes human systemic Lyme disease.

The results of this study suggest that HiMLST is a promising approach to perform epidemiological surveillance directly from infected biological material. In spite of the various information embedded in genetic data, ecological and medical data are required to obtain a complete picture of zoonotic pathosystems.

ACKNOWLEDGMENTS

We thank the staff of the GATC, especially Zélie Dubreucq and Marie-Elise Galoppini, for handling the sequencing process. This work was funded by the French National Institute for Agricultural Research (INRA) through the 'AIP Bioressources 2011' program. Maude Jacquot is supported by a PhD fellowship from the French Ministry of Higher Education and Research (MESR). Interactions between members of the INRA and the Pasteur Institute are funded by the 'Tick and Tick-Borne Diseases' (TMT) group, which is part of the Research Consortium on Biological Interactions (GdR REID). Field sampling was conducted with the financial support of the Conseil Régional d'Ile-de-France and the Office National des Forêts. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

- Akaike H. 1974. A new look at the statistical model identification. *Autom. Control. IEEE Trans.* 1974.
- Altschul S, Gish W, Miller W. 1990. Basic local alignment search tool. *J. Mol. Biol.*
- Benson D a, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res.* 41:D36–D42.
- Bivand R, Pebesma E, Gómez-Rubio V. 2013. *Applied spatial data analysis with R* 2nd edition. UseR! Series, Springer.
- Brisson D, Dykhuizen DE. 2006. A modest model explains the distribution and abundance of *Borrelia burgdorferi* strains. *Am. J. Trop. Med. Hyg.* 74:615–622.
- Brisson D, Dykhuizen DE. 2004. *ospC* diversity in *Borrelia burgdorferi*: different hosts are different niches. *Genetics* 168:713–722.
- Brisson D, Dykhuizen DE, Ostfeld RS. 2008. Conspicuous impacts of inconspicuous hosts on the Lyme disease epidemic. *Proc. R. Soc. B Biol. Sci.* 275:227–235.
- Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21:255–265.
- Clauset A, Newman MEJ, Moore C. 2004. Finding community structure in very large networks. *Phys. Rev. E, Stat. nonlinear, soft matter Phys.* 70.
- Cleaveland S, Laurenson MK, Taylor LH. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 356:991–9.
- Crowder CD, Matthews HE, Schutzer S, Rounds M, Luft BJ, Nolte O, Campbell SR, Phillipson C a, Li F, Sampath R, Ecker DJ, Eshoo MW. 2010. Genotypic variation and mixtures of Lyme *Borrelia* in *Ixodes* ticks from North America and Europe. *PLoS One* 5:e10650.
- Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Syst.*
- Dubska L, Literak I, Kocianova E, Taragelova V, Sychra O. 2009. Differential role of passerine birds in distribution of *Borrelia* spirochetes, based on data from ticks collected from birds during the postbreeding migration period in Central Europe. *Appl. Environ. Microbiol.* 75:596–602.
- Dykhuizen DE, Brisson D, Sandigursky S, Wormser GP, Nowakowski J, Nadelman RB, Schwartz I. 2008. Short report: The propensity of different *Borrelia burgdorferi* sensu stricto

genotypes to cause disseminated infections in humans. *Am. J. Trop. Med. Hyg.* 78:806–810.

Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* 52:696–704.

Guttman D, Wang P, Wang I-N, Bosler EMB, Luft BJ, Dykhuizen DE. 1996. Multiple infections of *Ixodes scapularis* ticks by *Borrelia burgdorferi* as revealed by single-strand conformation polymorphism analysis. *J. Clin. Microbiol.* 34:652–656.

Hanincová K, Taragelová V, Koci J, Schäfer SM, Hails R, Ullmann AJ, Piesman J, Labuda M, Kurtenbach K. 2003. Association of *Borrelia garinii* and *B. valaisiana* with songbirds in Slovakia. *Appl. Environ. Microbiol.* 69:2825–2830.

Haydon DT, Cleaveland S, Taylor LH, Laurenson MK. 2002. Identifying reservoirs of infection: a conceptual and practical challenge. *Emerg. Infect. Dis.* 8:1468–1473.

Herrmann C, Gern L, Voordouw MJ. 2013. Species co-occurrence patterns among Lyme borreliosis pathogens in the tick vector *Ixodes ricinus*. *Appl. Environ. Microbiol.* 79:7273–80.

Humair P, Douet V, Cadenas F, Schouis L, Van De Pol I, Gern L. 2007. Molecular identification of bloodmeal source in *Ixodes ricinus* ticks using 12S rDNA as a genetic marker. *J. Med. Entomol.* 44:869–880.

Humair P, Postic D, Wallich R, Gern L. 1998. An avian reservoir (*Turdus merula*) of the Lyme borreliosis spirochetes. *Zentralblatt für Bakteriologie*. 538:521–538.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254–267.

Jacquot M, Bisseux M, Abrial D, Marsot M, Ferquel E, Chapuis J-L, Vourc'h G, Bailly X. 2014. High-throughput sequence typing reveals genetic differentiation and host specialization among populations of the *Borrelia burgdorferi* species complex that infect rodents. *PLoS One* 9:e88581.

Jones KE, Patel NG, Levy M a, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008. Global trends in emerging infectious diseases. *Nature* 451:990–993.

Kurtenbach K, Peacey M, Rijpkema SG, Hoodless AN, Nuttall PA, Randolph SE. 1998. Differential transmission of the genospecies of *Borrelia burgdorferi* sensu lato by game birds and small rodents in England. *Appl. Environ. Microbiol.* 64:1169–1174.

Lassmann T, Sonnhammer ELL. 2005. Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6:298.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence

Chapitre 3

typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* 95:3140–5.

Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* 11:728–736.

Margos G, Gatewood AG, Aanensen DM, Hanincová K, Terekhova D, Vollmer S, Cornet M, Piesman J, Donaghy M, Bormane A, Hurn M, Feil EJ, Fish D, Casjens S, Wormser GP, Schwartz I, Kurtenbach K. 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. U. S. A.* 105:8730–5.

Margos G, Vollmer S, Ogden NH, Fish D. 2011. Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. *Infect. Genet. Evol.* 11:1545–1563.

Marsot M, Henry P-Y, Vourc'h G, Gasqui P, Ferquel E, Laignel J, Grysan M, Chapuis J-L. 2012. Which forest bird species are the main hosts of the tick, *Ixodes ricinus*, the vector of *Borrelia burgdorferi* sensu lato, during the breeding season? *Int. J. Parasitol.* 42:781–8.

Marsot M, Sigaud M, Chapuis JL, Ferquel E, Cornet M, Vourc'h G. 2011. Introduced Siberian chipmunks (*Tamias sibiricus barberi*) harbor more-diverse *Borrelia burgdorferi* sensu lato genospecies than native bank voles (*Myodes glareolus*). *Appl. Environ. Microbiol.* 77:5716–521.

Marsot M, Chapuis J-L, Gasqui P, Dozières A, Masségli S, Pisanu B, Ferquel E, Vourc'h G. 2013. Introduced Siberian chipmunks (*Tamias sibiricus barberi*) contribute more to lyme borreliosis risk than native reservoir rodents. *PLoS One* 8.

Olsén B, Jaenson T, Bergström S. 1995. Prevalence of *Borrelia burgdorferi* sensu lato-infected ticks on migrating birds. *Appl. Environ. Microbiol.* 61:3082–3087.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.

Pisanu B, Marsot M, Marmet J, Chapuis J-L, Réale D, Vourc'h G. 2010. Introduced Siberian chipmunks are more heavily infested by ixodid ticks than are native bank voles in a suburban forest in France. *Int. J. Parasitol.* 40:1277–1283.

Qiu W, Dykhuizen D, Acosta M, Luft B. 2002. Geographic uniformity of the Lyme disease spirochete (*Borrelia burgdorferi*) and its shared history with tick vector (*Ixodes scapularis*) in the northeastern United States. *Genetics* 160:833–849.

- Rauter C, Hartung T. 2005. Prevalence of *Borrelia burgdorferi* sensu lato genospecies in *Ixodes ricinus* ticks in Europe: a Metaanalysis. Appl. Environ. Microbiol. 71:7203–7216.
- Rudenko N, Golovchenko M, Hönig V, Mallátová N, Krbková L, Mikulásek P, Fedorova N, Belfiore NM, Grubhoffer L, Lane RS, Oliver JH. 2013. Detection of *Borrelia burgdorferi* sensu stricto *ospC* alleles associated with human lyme borreliosis worldwide in non-human-biting tick *Ixodes affinis* and rodent hosts in Southeastern United States. Appl. Environ. Microbiol. 79:1444–1453.
- Seinost G, Dykhuizen D. 1999. Four clones of *Borrelia burgdorferi* sensu stricto cause invasive infection in humans. Infect. Immun. 67:3518–3524.
- Stanek G, Reiter M. 2011. The expanding Lyme *Borrelia* complex—clinical significance of genomic species? Clin. Microbiol. Infect. 17:497–493.
- Swofford D. 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0 b10 2003.
- Tsao K, Bent SJ, Fish D. 2013. Identification of *Borrelia burgdorferi ospC* genotypes in host tissue and feeding ticks by terminal restriction fragment length polymorphisms. Appl. Environ. Microbiol. 79:958–964.
- Vaumourin E, Vourc'h G, Telfer S, Lambin X, Salih D, Seitzer U, Morand S, Charbonnel N, Vayssier-Taussat M, Gasqui P. 2014. To be or not to be associated: power study of four statistical modeling approaches to identify parasite associations in cross-sectional studies. Front. Cell. Infect. Microbiol. 4:1–11.
- Viana M, Mancy R, Biek R, Cleaveland S, Cross PC, Lloyd-Smith JO, Haydon DT. 2014. Assembling evidence for identifying reservoirs of infection. Trends Ecol. Evol. 29:270–279.
- Woolhouse MEJ, Gowtage-Sequeria S. 2005. Host range and emerging and reemerging pathogens. Emerg. Infect. Dis. 11:1842–7.
- Woolhouse MEJ, Taylor LH, Haydon DT. 2001. Population biology of multihost pathogens. Science (80-.). 292:1109–1112.
- Wright S. 1965. The interpretation of population structure by F-Statistics with special regard to systems of mating. Soc. Study Evol. 19:395–420.

Données supplémentaires de l'article 3

Table S1 : Number of ticks, chipmunks, and wood mice sampled infected by each group of genotypes for the two loci

Genes	Groups of genotypes	Number of nymphs	Number of chipmunks	Number of bank voles	Number of wood mice
<i>rplB</i>	G1	93	45	1	0
<i>rplB</i>	G2	5	2	0	0
<i>rplB</i>	G3	13	1	90	4
<i>rplB</i>	G4	121	51	1	4
<i>rplB</i>	G5	41	2	1	0
<i>rplB</i>	G6	6	1	0	0
<i>rplB</i>	G7	49	0	0	0
<i>rplB</i>	G8	27	1	0	0
<i>rplB</i>	G9	1	0	0	0
<i>rplB</i>	G10	4	0	0	0
<i>rplB</i>	G11	2	0	0	0
<i>rplB</i>	G12	1	0	0	0
<i>rplB</i>	G13	1	0	0	0
<i>rplB</i>	G14	7	0	0	0
<i>rplB</i>	G15	1	0	0	0
<i>rplB</i>	G16	1	0	0	0
<i>rplB</i>	G17	48	0	0	0
<i>ospC</i>	G1	35	24	3	0
<i>ospC</i>	G2	31	3	0	1
<i>ospC</i>	G3	12	0	55	0
<i>ospC</i>	G4	22	4	0	0

Chapitre 3

<i>ospC</i>	G5	32	2	1	0
<i>ospC</i>	G6	45	9	0	0
<i>ospC</i>	G7	64	4	0	0
<i>ospC</i>	G8	36	1	37	1
<i>ospC</i>	G9	22	2	0	0
<i>ospC</i>	G10	52	14	0	1
<i>ospC</i>	G11	43	22	0	0
<i>ospC</i>	G12	11	1	0	0
<i>ospC</i>	G13	0	5	1	0
<i>ospC</i>	G14	38	17	0	0
<i>ospC</i>	G15	20	0	0	0
<i>ospC</i>	G16	9	0	0	0
<i>ospC</i>	G17	5	1	0	0
<i>ospC</i>	G18	3	0	0	0
<i>ospC</i>	G19	8	0	0	0
<i>ospC</i>	G20	14	0	0	0
<i>ospC</i>	G21	38	1	0	0
<i>ospC</i>	G22	1	0	0	0
<i>ospC</i>	G23	11	0	0	0
<i>ospC</i>	G24	3	0	0	0
<i>ospC</i>	G25	8	0	0	0
<i>ospC</i>	G26	2	0	0	0
<i>ospC</i>	G27	5	0	0	0
<i>ospC</i>	G28	16	0	0	0

DISCUSSION ET PERSPECTIVES

Les bactéries pathogènes du complexe d'espèces *B. burgdorferi* s.l. circulent chez différentes espèces de vecteurs en Europe et aux États-Unis [Stanek *et al.*, 2012] et elles sont capables d'infecter une grande diversité d'hôtes. Lorsque l'on s'intéresse à la diversité génétique de la bactérie, on se rend compte que de nombreux groupes génétiques distincts sont maintenus au sein du complexe [Margos *et al.*, 2011]. C'est dans ce contexte que s'est inscrite ma thèse, au cours de laquelle nous avons mené différentes études, afin d'apporter de nouvelles clés pour mieux comprendre comment les bactéries de *B. burgdorferi* s.l. circulent entre leurs vecteurs et leurs hôtes. En effet, nous avons cherché à définir quels étaient les scénarios évolutifs et les processus épidémiologiques les plus cohérents pour expliquer la distribution de la diversité génétique que l'on peut observer. Pour cela, nous nous sommes intéressés aux processus de diversification *via* l'étude de la recombinaison homologue, des migrations entre les populations du complexe *B. burgdorferi* s.l., des démographies et des interactions moléculaires entre les pathogènes et le système immunitaire des hôtes (pressions de sélection). Ces différents points vont être replacés brièvement dans leur contexte, puis les principaux résultats que nous avons obtenus seront discutés et mis en perspective.

1. Isolement génétique des différentes génoespèces de *B. burgdorferi* s.l. et recombinaison

Le développement des outils de génétique et de génomique autorise l'étude des processus évolutifs qui façonnent la diversité génétique des pathogènes. Dans le cadre de l'étude du complexe *B. burgdorferi* s.l., des chercheurs américains ont notamment mené une étude de génomique comparative avec l'objectif de comprendre comment la diversité au sein des populations apparaît et est maintenue en présence de possibles recombinaisons fréquentes [Haven *et al.*, 2011]. Leur étude est basée sur l'analyse de 23 génomes de souches principalement de l'espèce *B. burgdorferi* s.s. sélectionnées pour capturer un maximum de diversité génétique [Schutzer *et al.*, 2011], compromettant de ce fait les estimations de taux de recombinaison homologue (**Article 1**). Malgré un biais qui peut être dû à la mise en culture des souches de *Borrelia*, nous avons utilisé pour notre part un **échantillonnage local et aléatoire** qui permet de représenter les fréquences des différents génotypes des espèces dans les tiques à l'affût des sites d'échantillonnage près de Munster et Guebwiller. De plus, notre échantillon comporte, pour à minima deux espèces (*B. burgdorferi* s.s. et *B. garinii*), un nombre de souches suffisant pour obtenir un pouvoir statistique adapté pour les analyses réalisées par la suite.

Discussion et perspectives

Les approches phylogénétiques que nous avons menées (**Figure 2 Article 1**) montrent que le complexe d'espèces *B. burgdorferi* s.l. est composé de différents **groupes génétiques** qui sont clairement **isolés** les uns des autres. Ces résultats sont cohérents avec les frontières génétiques robustes observées entre les génoespèces du complexe à partir des analyses de MLST [Margos *et al.*, 2011]. Le niveau très fort de différenciation génétique et les valeurs de déséquilibre de liaison en moyenne plus élevées à l'échelle inter-spécifique qu'à l'échelle intra-spécifique renforcent ce point de vue. De plus, les résultats obtenus à partir du modèle de coalescent que nous avons construit permettent d'estimer que les événements de recombinaison homologue sont environ 50 fois plus fréquents au sein des génoespèces qu'entre les génoespèces. Nous avons estimé un rapport de taux de recombinaison/taux de mutation de 1.7 ce qui est, bien que plus faible, cohérent avec ce qui avait été obtenu par Haven et collaborateurs [2011] pour les souches de *B. burgdorferi* s.s..

Seules les souches de *B. bavariensis* et de *B. finlandensis* ont une quantité moyenne de divergence de moins de 5 % avec les souches des espèces de *B. garinii* et *B. burgdorferi* s.s. respectivement, ce qui pose question quant à leur **réelle appartenance à une autre espèce** (**Article 1**). Une étude de génomique comparative se basant sur les 23 génomes préalablement séquencés a fourni des résultats du même ordre [Mongodin *et al.*, 2013]. Nous avons avancé deux hypothèses majeures pour expliquer ces pourcentages de divergence faible (**Article 1**). Tout d'abord, les taux de recombinaison homologue entre les souches de ces couples de génoespèces peuvent être importants et prévenir la diversification. Cependant, cette hypothèse est aisément remise en question puisque, bien que nous n'ayons pas d'information concernant la séparation entre *B. finlandensis* et *B. burgdorferi* s.s., la divergence entre *B. bavariensis* et *B. garinii* est associée à un changement de spectre d'hôte [Margos *et al.*, 2009] suggérant un manque de contact physique empêchant l'homogénéisation des lignées par la recombinaison homologue et autorisant ainsi l'accumulation de mutations par dérive ou sélection. La seconde hypothèse, plus probable, voudrait que le temps depuis lequel les lignées divergent ne soit pas encore suffisant pour permettre leur séparation complète.

En relation directe avec cette dernière hypothèse, il serait intéressant d'essayer de prévoir à quelle fréquence de nouveaux groupes génomiques de *Borrelia* sont capables d'émerger. Ceci permettrait d'établir une échelle d'ordonnancement, voire chronologique, d'évolution et de lien des génoespèces entre elles et constituerait une avancée importante dans

la **connaissance de l'histoire et des modalités évolutives** des différentes espèces du complexe. Pour cela, on peut passer par l'étude des événements passés.

Une étude suggère que c'est un phénomène d'**explosion radiative** qui est à l'origine de la diversification des différentes espèces de *B. burgdorferi* s.l. et que les taux de diversification au sein du complexe sont aujourd'hui plus faibles [Morlon *et al.*, 2012]. Cependant les analyses ont été basées sur des phylogénies construites, d'une part à partir de données de séquences multilocus qui peuvent ne pas porter suffisamment d'information aux vues de ce que l'on observe à l'échelle du génome, et d'autre part, à partir de séquences génomiques ne représentant qu'une faible part de la diversité réelle du complexe. Le séquençage de génomes complets d'un grand nombre de souches échantillonnées aléatoirement permettrait d'obtenir une meilleure image des taux de diversification.

En faisant des hypothèses sur le taux de mutation, notre modèle de coalescent est susceptible de donner des informations sur la date des événements de divergence. En effet, il a été estimé que le temps depuis lequel les populations de *B. burgdorferi* s.s. et *B. garinii* divergent l'une de l'autre est d'environ 500 000 générations (**Article 1**). Une autre approche faisant des hypothèses sur le taux de mutation et la distribution des valeurs sélectives des allèles *ospC* ont aussi été utilisées pour dater à environ 6 500 000 générations l'introduction, depuis l'Europe, de *B. burgdorferi* s.s. aux États-Unis [Mongodin *et al.*, 2013]. Les valeurs obtenues dans les deux cas sont fortement influencées par les hypothèses des modèles et ne semblent *a priori* pas cohérentes. Ces résultats rappellent que, de façon générale, pour dater correctement les événements de divergence entre des espèces ou de colonisation, il faudrait **calibrer l'horloge moléculaire** des différentes espèces du complexe. Par exemple, le logiciel BEAST qui permet ce calibrage [Drummond and Rambaut, 2007] a été utilisé sur 62 gènes de *Mycoplama mycoides*, un agent responsable de pneumonies chez les bovins, pour reconstruire l'histoire évolutive du pathogène, estimer la date de son émergence et déterminer l'origine de certaines épidémies [Dupuy *et al.*, 2012]. De façon similaire, cette approche pourrait être utilisée sur des données de génomes de souches de *Borrelia* échantillonnées sur de nombreuses années en Europe et aux États-Unis pour dater plus précisément l'introduction de *B. burgdorferi* s.l. aux États-Unis et les événements de diversification des différentes espèces en Europe.

Les processus de diversification peuvent être associés à l'établissement de populations dans de nouvelles niches écologiques auxquelles elles doivent s'adapter. L'identification de ce

Discussion et perspectives

qui caractérise ces niches, est donc une question d'intérêt.

2. Origine de l'émergence de nouveaux groupes de *Borrelia* et adaptation à un nouvel environnement

Les trois unités de répllication les plus conservées du complexe d'espèces *B. burgdorferi* s.l. ont été balayées par diverses statistiques descriptives (**Article 1**) afin de rechercher les régions du génome pouvant être soumises à des pressions de sélection, synonyme d'**adaptation de la bactérie à l'environnement**. Ce sont de potentielles adaptations à différentes niches écologiques qui ont pu conduire à la diversification des différentes espèces de *Borrelia*. De façon congruente avec d'autres études, ces analyses ont permis de souligner i) la sur-représentation des gènes impliqués dans la production de lipoprotéines dans ces régions sous pressions de sélection [Brisson *et al.*, 2012 ; Mongodin *et al.*, 2013], en raison de leur interaction avec l'environnement, mais surtout ii) la singularité de la région du plasmide cp26 proche du gène *ospC* qui est soumise à des pressions de sélection diversifiantes et/ou fréquence-dépendantes négatives [Brisson and Dykhuizen 2004 ; Qiu *et al.*, 1997] (**Article 2**).

Il est probable que de nombreux autres gènes responsables de l'adaptation de *B. burgdorferi* s.l. à l'environnement soient présents sur les **autres plasmides de la bactérie**. Pour étudier ces plasmides, un nouveau séquençage, avec une couverture beaucoup plus importante, serait nécessaire en raison des difficultés que nous avons rencontrées pour assembler et aligner leurs séquences pour notre étude (**Article 1**). De plus, les souches dont les génomes ont été séquencés ont été isolées à partir de tiques à l'affût dont on ne sait pas sur quel hôte elles se sont gorgées. De ce fait, on n'a pas *a priori* sur les données et il est difficile de conclure de façon certaine quant aux facteurs qui exercent les pressions de sélection. Ainsi, la mise en œuvre d'un échantillonnage plus structuré passant, par exemple, par l'isolement de souches à partir d'hôtes différents ou de vecteurs différents, permettrait d'obtenir davantage d'information concernant l'impact des différents facteurs qui peuvent être à l'origine de l'émergence de différents groupes génétiques.

Par des approches plus expérimentales, on peut vouloir essayer de **prédire** quelles pourraient être les **mutations et/ou recombinaisons** les plus **susceptibles d'apparaître** dans le futur en raison d'une meilleure adaptation et donc d'un fort potentiel épidémique. Par

exemple, très récemment une étude a cherché à mettre en place une méthode permettant de prédire les mutations du virus responsable du chikungunya les plus susceptibles d'émerger à court terme et présentant un fort potentiel épidémiologique [Stapleford *et al.*, 2014]. Basée sur des protocoles expérimentaux, faisant intervenir des cycles de transmission complets entre vecteur et hôtes du virus, combinés au séquençage de séquences génomiques du virus, l'étude a permis de mettre en avant le fort potentiel émergent de deux nouvelles souches mutées. Des analyses *in vitro* suggèrent que les avantages sélectifs de ces deux mutations pourraient être liés à une capacité accrue du virus à fusionner avec les membranes de cellules cibles de l'infection et à une meilleure stabilité thermique dans les milieux extra-cellulaires comme la salive ou le sang des mammifères. Ce type de méthodes ouvre des perspectives nouvelles, tant pour la surveillance des épidémies que pour les recherches vaccinales menées sur de nombreux virus pathogènes mais leur applicabilité pour les bactéries sera à vérifier, notamment puisque les génomes des bactéries évoluent moins rapidement que ceux des virus [Gago *et al.*, 2009].

L'adaptation des génoespèces de *B. burgdorferi* s.l. à différentes niches écologiques peut conduire, ou arriver de façon concomitante, à des modalités de circulation entre les hôtes et les vecteurs différentes en fonction des populations de *Borrelia* impliquées. Ce sont ces modalités qui ont été étudiées dans la suite de mon projet de thèse.

3. Circulation des populations *B. burgdorferi* s.l. entre les hôtes et les vecteurs

3.1. Dynamiques de circulation hôte-spécifiques des différentes génoespèces de *Borrelia*

Des études précédentes, basées sur des données de MLST, ont montré que les tailles de populations et les migrations potentielles des hôtes réservoirs peuvent avoir un impact sur les patrons de diversité de *B. burgdorferi* s.l. [Vollmer *et al.*, 2011]. Cependant, ces résultats ont une dimension continentale. Nos travaux, à l'échelle des deux sites de Munster et Guebwiller, apportent donc de nouvelles réponses concernant l'influence de ces processus à une échelle locale (**Article 1**). Le développement d'un modèle de phylodynamique nous a permis de comprendre comment les **processus épidémiologiques et d'évolution** interagissent pour façonner les phylogénies des génoespèces de *Borrelia* étudiées qui circulent dans les deux sites d'échantillonnage entre les populations tiques et hôtes. Nous avons pu conclure que les différences, à l'échelle génomique, de niveaux de diversité génétique et de niveaux de

Discussion et perspectives

différenciation entre les deux localités des deux espèces, *B. burgdorferi* s.s. et *B. garinii*, peuvent s'expliquer par des dynamiques de circulation différentes dues à des spectres d'hôtes d'infection différents. En effet, la population d'hôtes infectée par *B. burgdorferi* s.s. semble avoir une taille plus faible et un renouvellement plus important que celle infectée par *B. garinii*. Les individus de la population d'hôtes infectée par *B. garinii* semblent migrer, d'un site à l'autre, plus fréquemment que ceux de la population d'hôtes infectée par *B. burgdorferi* s.s.. Ces résultats sont cohérents avec les observations faites en Europe et qui montrent que *B. garinii* infecte plutôt des oiseaux, et avec l'hypothèse que dans nos sites d'échantillonnage, *B. burgdorferi* s.s. infecte plutôt des petits rongeurs. Ils seraient également la première preuve à l'échelle génomique de l'influence du spectre d'hôte sur l'évolution et la dynamique de la bactérie.

En Europe, les associations, à l'échelle des espèces de *B. burgdorferi* s.l., aux différents types d'hôtes sont plutôt bien décrites dans la littérature [Kurtenbach *et al.*, 2002 ; Kurtenbach *et al.*, 2006 ; Margos *et al.*, 2011]. En revanche, seulement peu d'information concernant ce qu'il se passe à l'échelle intra-spécifique était jusque-là disponible. L'analyse de la variabilité moléculaire que nous avons menée sur les séquences de *Borrelia* infectant les tamias et les campagnols de Sénart, montre qu'il existe un **niveau de différenciation fort des génotypes en fonction des espèces d'hôte**, y compris à l'échelle intra-spécifique, aux deux marqueurs que nous avons utilisés (**Article 2**). Ces résultats étaient attendus pour le gène *ospC* en raison de ses fonctions. En revanche, la différenciation des génotypes du gène *rplB* suggère, quant à elle, que certaines espèces d'hôtes conduisent à l'émergence de lignées qui leur sont spécifiquement associées. Bien que les travaux de génomique montrent que la diversité au sein des groupes génétiques semble distribuée largement à l'échelle géographique (**Article 1**), à Sénart nous avons pu observer des groupes qui, pour ce que l'on en sait, sont site-spécifiques comme le génotype *rplB* G3 par exemple (**Articles 2 et 3**). De plus, nous avons eu l'occasion de caractériser les génotypes *rplB* et *ospC* d'une quinzaine de campagnols roussâtres provenant d'autres forêts en France et en Espagne et nous n'avons pas retrouvé les génotypes identifiés chez les campagnols échantillonnés à Sénart (données non montrées). Ces différents résultats posent questions concernant la façon dont se créent localement les associations entre les espèces d'hôtes et les bactéries et sur les potentielles adaptations qui peuvent se mettre en place lors de la création de ces associations. D'autres études seront à effectuer afin d'apporter davantage de réponses à ces questions. Il serait par exemple

intéressant de caractériser les *B. burgdorferi* s.l. infectant différentes espèces d'hôtes dans plusieurs sites et d'étudier de façon précise les mutations, synonymes et non synonymes, qui impactent leurs séquences.

Bien que de nouvelles études, avec plus de loci, seraient nécessaires pour mieux comprendre les possibles émergences de lignées spécifiques aux hôtes, les résultats acquis par notre étude pourraient avoir un impact important dans la compréhension de la **circulation de la maladie** en Europe mais également, **aux États-Unis** ou une seule géoespèce de *Borrelia*, pathogène pour l'homme, est majoritairement retrouvée [Kurtenbach *et al.*, 2006]. En effet, un nombre restreint d'espèces ou de populations d'hôtes pourrait être responsable de la circulation des génotypes qui provoquent la maladie de Lyme chez l'homme et leur identification permettrait un bien meilleur contrôle de la maladie. Contrôler les populations de vertébrés peut s'avérer complexe et lourd de conséquences ; en revanche, la diffusion de consignes de comportements adaptés pour l'homme dans les zones à risques, c'est à dire où il y a présence des hôtes porteurs des génotypes qui posent problème, permettrait sans aucun doute, de réduire le nombre de nouveaux cas.

Les associations entre les espèces d'hôtes et les génotypes de *B. burgdorferi* s.l. observées pourraient-être liées à l'établissement possible de groupes de **tiques spécialisées à différents types d'hôtes** [Kempf *et al.*, 2011]. Si comme le suggère cette étude, dans des populations locales, le choix de l'hôte par la tique *I. ricinus* est non aléatoire et le résultat de tiques spécialisées à différents types d'hôtes, les différentes lignées de *B. burgdorferi* s.l. pourraient circuler dans un système hôte-vecteur fermé. Cependant, on peut imaginer qu'à long terme les différentes populations de *B. burgdorferi* s.l., ne pouvant plus être homogénéisées *via* la recombinaison et la migration, divergeraient suffisamment pour former des espèces différentes. Des études supplémentaires mettant l'évolution des pathogènes en parallèle avec l'écologie et l'évolution des vecteurs, devront être réalisées pour permettre d'apporter d'avantage de réponses à ce sujet-là.

Le modèle de phylodynamique simple que nous avons développé a permis d'apporter des éléments de compréhension concernant la maintenance de différentes lignées dans des populations d'hôtes de tailles réalistes (**Article 1**). Cependant, le développement de **modèles phylodynamiques incorporant des contraintes évolutives** comme la spécialisation

Discussion et perspectives

adaptative ou la sélection fréquence-dépendante, par exemple, permettrait sans doute une meilleure compréhension d'ensemble des pathosystèmes faisant intervenir des pathogènes multi-hôtes (Figure 18).

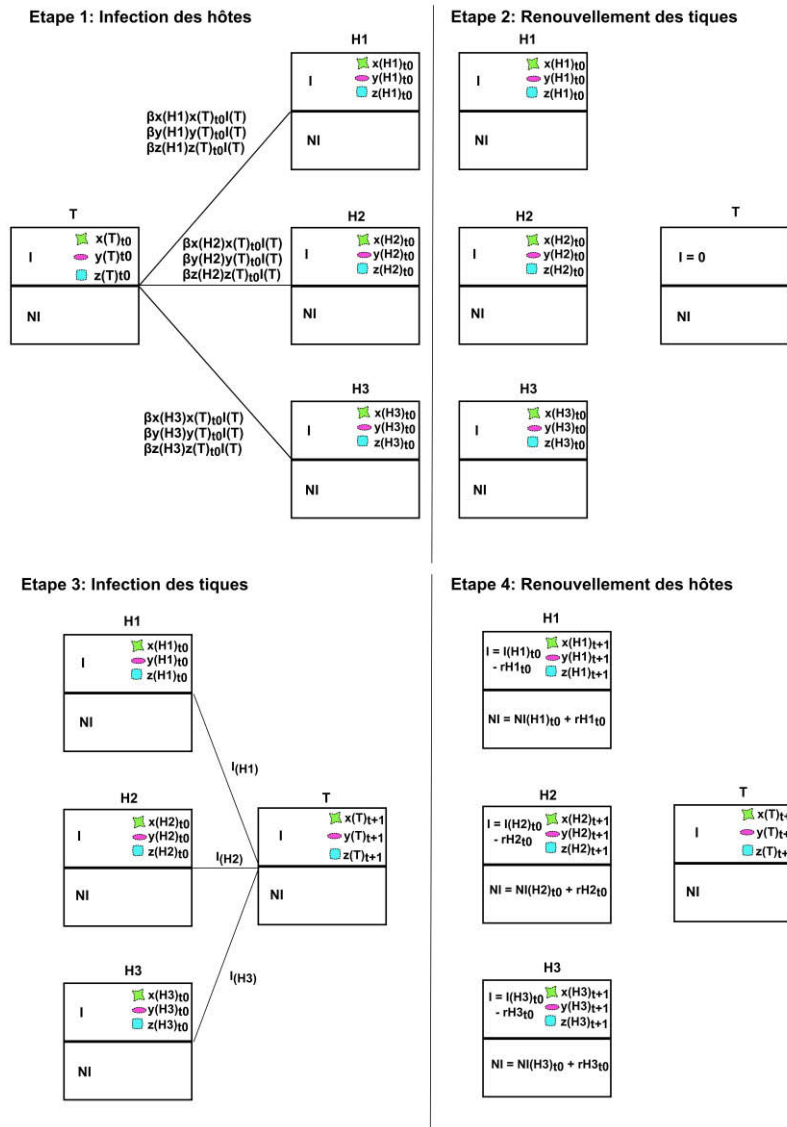


Figure 18 : Schématisation d'un modèle phylodynamique qui prendrait en compte la spécialisation adaptative des *Borrelia*.

Cas d'un modèle à transmission densité dépendante. Chaque logo de couleur représente une lignée de *Borrelia* différente. À l'étape 1, les tiques infectées (I) vont se nourrir sur les différents hôtes, qui peuvent être déjà infectés ou non (NI), et leur transmettre les bactéries en fonction des taux de transmission de chaque lignée pour l'hôte donné (prise en compte des spécialisations possibles). À l'étape 2, toutes les nymphes sont remplacées par des larves qui sont non-infectées. À l'étape 3, les larves vont alors se nourrir sur les hôtes et acquérir la bactérie avec un taux de transmission uniquement dépendant de la densité d'hôte. À l'étape 4, les larves sont métamorphosées en nymphes, et une proportion r des hôtes est renouvelée (mort/naissance), réduisant le nombre d'infectés. On peut alors repartir à l'étape 1 et recommencer ces étapes autant d'itérations qu'il est nécessaire pour obtenir un équilibre.

En raison de nombreuses variations des taux de transmission, des charges en tiques ou de comportements par exemple, les différentes espèces d'hôtes d'un pathosystème ne contribuent pas de la même façon au risque lié à la maladie. Nous avons donc étudié les contributions des rongeurs échantillonnés à Sénart.

3.2. Contribution des différentes espèces d'hôtes

Malgré l'importance que les co-infections peuvent avoir, les méthodes classiques d'épidémiologie moléculaire ne permettent pas d'identifier proprement les différents génotypes qui peuvent infecter simultanément un même individu. Dans ce contexte, il est intéressant de développer des méthodologies qui permettent d'obtenir une information précise d'ordre qualitative et quantitative des niveaux de co-infections des individus. Nous avons donc réalisé un séquençage haut débit des amplifiats de deux marqueurs de *Borrelia* infectant trois espèces de rongeurs échantillonnés en forêt de Sénart [Marsot *et al.*, 2011 ; Marsot *et al.*, 2013]. Ceci a permis de **caractériser les profils d'infection** (mono et co-infections) de nombreux individus tamias de Sibérie, campagnols roussâtres et mulots sylvestres aux deux loci séquencés. Bien qu'une majorité d'individus hôtes étaient mono-infectés aux deux loci, ce séquençage haut-débit a permis de détecter et caractériser jusqu'à 5 génotypes différents du gène *rplB* chez un campagnol et 5 génotypes du gène *ospC* chez un individu tamia ([Figure 19](#)). L'approche utilisée peut être améliorée en augmentant le nombre de marqueurs utilisés ce qui se traduirait par une augmentation du nombre de polymorphismes sur lesquels on travaille et donc du degré de précision pour l'établissement des relations de parenté entre les génotypes séquencés [Maiden *et al.*, 2013]. Cependant, la HiMLST est une méthode efficace, de qualité et à bas coûts pour caractériser les espèces bactériennes [Boers *et al.*, 2012]. Elle permet également d'aller plus loin dans l'analyse des profils d'infection d'individus. Par exemple, il nous a été possible d'assigner les hôtes et les tiques étudiés (**Article 3**) à des cycles, site-spécifiques, potentiels de transmission de *B. burgdorferi* s.l. faisant émerger de nouvelles questions concernant les dynamiques d'infection des individus qui pourraient éventuellement faire intervenir des processus de compétition intra-individus et conduire à des niveaux de diversité et de co-infections variables en fonction des espèces d'hôte considérées.

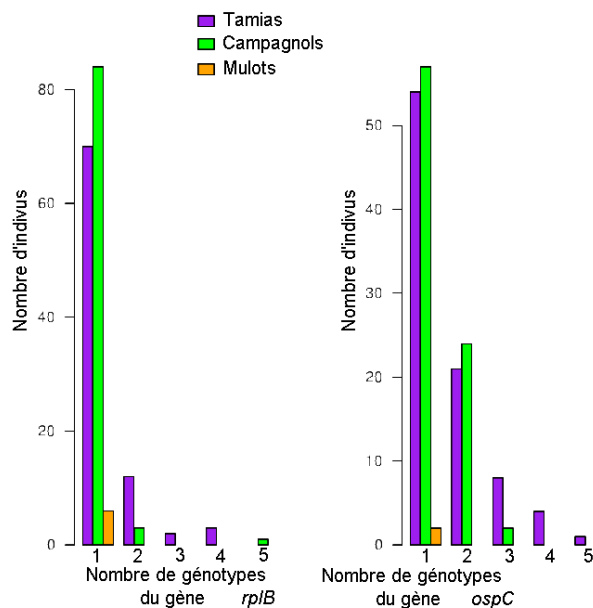


Figure 19 : Distribution du nombre de génotypes détectés chez les individus hôtes aux loci *rplB* et *ospC*

En effet, grâce à cette méthodologie, nous avons pu observer que **la diversité de génotypes et les niveaux de co-infections étaient plus forts chez le tamia** de Sibérie que chez les campagnols roussâtres. Nous avons émis différentes hypothèses, non exclusives, pour expliquer ce résultat important (**Article 2**). Tout d'abord, les tamias ont une charge en tiques plus importante que les campagnols roussâtres [Pisanu *et al.*, 2010], pouvant conduire à une exposition plus importante de cette espèce hôte à la bactérie et donc à un risque d'infection et co-infection plus fort. Deuxièmement, comme les tamias de Sibérie ont colonisé la forêt de Sénart entre les années 1970 et 1990 [Chapuis *et al.*, 2011], il est possible qu'ils aient alloué un maximum de ressources à la reproduction pour garantir le succès de leur établissement, au détriment de la réponse immunitaire [Lee and Klasing, 2004]. Ceci aurait pour conséquence une plus forte susceptibilité de l'espèce à l'infection par les *Borrelia*. Enfin, il est possible qu'un plus grand nombre de génotypes soient capables d'infecter les tamias que les campagnols. Ces résultats confirment ce qui avait été fortement suspecté suite aux travaux de thèse de Maud Marsot [Marsot, 2011] concernant le rôle important que jouent les tamias dans la circulation de *B. burgdorferi* s.l. et leur forte contribution au risque acarologique, c'est à dire lié à l'homme.

Les différents modèles développés pour étudier la contribution des hôtes à la circulation de la bactérie prennent le plus souvent en compte des paramètres

épidémiologiques tels que les taux de transmission ou bien des informations concernant la densité et l'abondance des hôtes [Randolph and Craine 1995 ; Brisson *et al.*, 2008]. Concernant les génotypes identifiés à Sénart, nous n'avons pas ces informations disponibles. Nous avons donc développé un nouveau **modèle statistique pour estimer les contributions relatives des hôtes** à partir des fréquences des génotypes de *B. burgdorferi* s.l., identifiés pour les deux marqueurs, dans les hôtes échantillonnés et dans les tiques à l'affût (**Article 3**). Ce modèle estime que les contributions du réservoir comprenant le tamia de Sibérie et du réservoir comprenant toutes les autres espèces d'hôtes (celles qui n'ont pas été échantillonnées) à la circulation de la bactérie sont très importantes tandis que celle du campagnol roussâtre a été estimée comme étant très faible. De nouvelles recherches sont nécessaires afin de déterminer si d'autres espèces d'hôtes, que nous n'avons pas échantillonnées, peuvent être infectés par les mêmes génotypes que le tamia afin de caractériser définitivement et de façon plus précise le ou les cycle(s) de transmission de ces génotypes. Cependant, les résultats obtenus soulignent le rôle important joué par les tamias de Sibérie, achetés dans les jardinerie et lâchés dans la nature par leurs propriétaires. Il apparaît alors important de sensibiliser les acheteurs de tamias et de contrôler leur population pour ne pas augmenter le risque lié à la maladie dans d'autres localités où le tamia pourrait s'installer suite à des lâchers.

Le modèle que nous avons développé est généralisable aux autres maladies multi-hôtes à transmission vectorielle, mais il est important de prendre en compte que plus le nombre d'hôtes pour la maladie est important, plus l'effort d'échantillonnage doit être conséquent pour que la part de la catégorie X (hôtes non échantillonnés) à la diversité observée chez le vecteur ne soit pas trop importante. En effet, lorsque l'on n'a pas *a priori*, il est compliqué d'émettre des hypothèses concernant les hôtes de la catégorie X, alors qu'elle peut avoir une forte contribution. L'utilisation de méthodes d'**identification de l'espèce d'hôte sur laquelle la tique a pris son dernier repas sanguin** permettrait sans doute de formuler des hypothèses quant à la composition de cette catégorie X, mais elles sont très difficiles à mettre au point en raison de la digestion de l'ADN des vertébrés dans l'estomac des tiques et le long délai qu'il peut y avoir entre la prise du repas par la tique et l'analyse en laboratoire [Kent, 2009]. Ces méthodes ont donc une faible efficacité [Humair *et al.*, 2007 ; Cadenas *et al.*, 2007] et peuvent conduire à des résultats peu fiables ; il a en effet été identifié des repas sur des espèces d'hôtes qui n'étaient pas censées être présentes dans les sites d'étude

Discussion et perspectives

[Marsot, 2011]. Les résultats également obtenus pour les tiques de notre étude à Sénart, n'ont, à ce jour, pas été concluants (Figure 20), et bien qu'en parallèle de ma thèse, une autre méthode, basée sur du séquençage haut-débit, a été développée pour permettre l'identification de la dernière espèce d'hôte sur laquelle s'est gorgée une tique [Masseglia *et al.*, données non publiées], de nouveaux efforts sont encore à faire pour fournir des résultats plus fiables et d'un rendement supérieur.

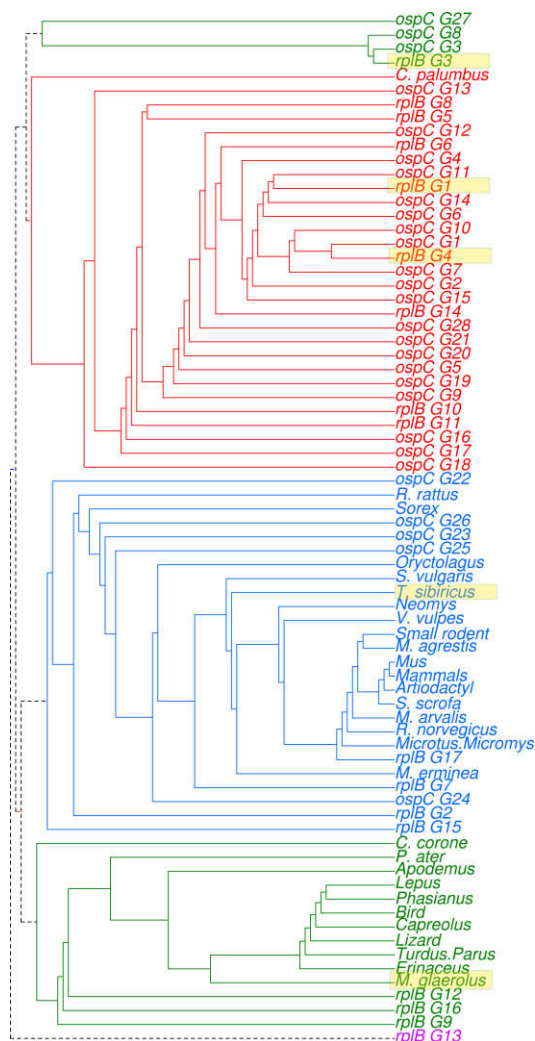


Figure 20 : Co-occurrence des génotypes identifiés chez les tiques et les hôtes et des identifications de la dernière espèce d'hôte sur laquelle se sont gorgées les tiques

Dendrogramme des génotypes identifiés par séquençage des deux marqueurs de *Borrelia* (*rplB* GX, *ospC* GX) et des espèces d'hôtes identifiées par la méthode d'identification des repas sanguins chez les tiques échantillonnées à Sénart. Ce dendrogramme est la représentation graphique des groupes de génotypes les plus fréquemment retrouvés chez les mêmes individus (une couleur différente) (fast.greedy sous R). Les génotypes *rplB* G1, *rplB* G4, *ospC* G1, *ospC* G11 et *ospC* G16 (spécifiques des tamias) et les génotypes *rplB* G3, *ospC* G3, *ospC* G8 (spécifiques des campagnols) (Article 2) ne se groupent pas avec, respectivement, les espèces *Tamias sibiricus* et *Myodes glareolus* de la méthode d'identification de repas sanguin, suggérant qu'il n'y a pas d'homogénéité des prédictions fournies par les deux méthodes.

Parmi les populations d'hôtes qui contribuent au réservoir de la maladie, il pourrait être intéressant de compléter l'information des contributions par l'identification des populations et/ou communautés qui permettent la maintenance ou non du pathogène [Haydon *et al.*, 2002]. Pour cela, il serait sans doute possible de travailler avec un modèle dynamique à compartiments, pour lequel on regarderai l'évolution du ou des compartiment(s) d'hôtes infectés en fonction de la présence ou non d'une espèce d'hôtes donnée.

Enfin, pour mieux contrôler la maladie, il sera important d'identifier les espèces d'hôtes chez lesquelles circulent, les lignées de *B. burgdorferi* s.l. qui infectent l'homme. Ceci permettra, éventuellement, de mettre en place des mesures de prévention pour limiter le contact des hommes avec les tiques qui auraient pu se nourrir sur ces hôtes à risque(s). De plus, les génotypes d'*ospC* de ces lignées pourront être la cible prioritaire pour le développement de traitements ou vaccins.

4. Conclusion

Les résultats obtenus pendant ma thèse contribuent à la connaissance des scénarios évolutifs et des processus épidémiologiques permettant d'expliquer la distribution de la diversité génétique de *B. burgdorferi* s.l.. Ce fut l'occasion de démontrer l'applicabilité et l'intérêt d'utiliser des approches de génomique comparative et de séquençage haut-débit de marqueurs dans le cadre de l'étude, d'une part, des processus évolutifs et d'autre part, des processus épidémiologiques, qui façonnent la diversité de pathogènes. Nous avons notamment montré que l'étude des co-infections des individus par différents génotypes permettait d'acquérir de plus amples informations concernant la dynamique de transmission des pathogènes et que ces dynamiques peuvent largement être étudiées en modélisant de façon plus ou moins complexe les systèmes étudiés. Des approches similaires peuvent être transposées pour l'étude d'autres maladies infectieuses.

Dans le futur, les études faisant appel à des modèles de phylodynamique permettront sans doute d'aborder de façon plus précise les mécanismes biologiques sous-jacents à l'évolution de différentes espèces de pathogènes. Ces informations sont cruciales pour permettre l'élaboration de stratégies efficaces dans le cadre de la lutte contre les maladies humaines et animales importantes, y compris les zoonoses comme la borréliose de Lyme. Par exemple, le développement des vaccins saisonniers contre le virus de la grippe est

Discussion et perspectives

principalement basé sur la prédiction de l'évolution du pathogène. Dans ce cas-là, une compréhension à une échelle phylodynamique de comment le virus évolue et circule dans les populations d'hôtes est une approche pertinente. Dans le cadre des pathosystèmes plus complexes faisant intervenir hôtes, vecteurs et surtout des pathogènes qui évoluent plus lentement que les virus, l'applicabilité des approches phylodynamiques reste encore à confirmer mais les résultats que nous avons obtenus (**Article 1**) sont prometteurs et doivent encourager d'avantage d'efforts dans cette direction.

De façon complémentaire, la mise en place d'études d'évolution expérimentale, combinant expérimentation en laboratoire et séquençage haut-débit, pourraient ouvrir de nouvelles perspectives pour comprendre le fonctionnement des systèmes biologiques et émettre des hypothèses sur le devenir de ces systèmes. En effet, on peut espérer être capable de prédire les comportements des maladies infectieuses en réponse à leur introduction dans un nouvel environnement ou aux changements climatiques par exemple. Il s'agira de comprendre comment certaines mutations peuvent changer les propriétés fonctionnelles des protéines et comment elles pourraient se propager dans les populations d'hôtes et de vecteurs sous l'effet de la sélection et de la dérive.

Ces deux approches complémentaires ne seront que plus efficaces si l'on est capable de spécifier des données *a priori* pour les modèles construits. En effet, ceci permettrait de mettre, plus efficacement, l'évolution des pathogènes en parallèle avec l'écologie et l'évolution de leurs hôtes et de leurs vecteurs afin d'obtenir une représentation globale du système que l'on étudie. Aujourd'hui, il est encore très difficile de mettre en place ce type d'approches *de novo*, et une alternative pourrait éventuellement provenir de la réutilisation de métadonnées déjà disponibles dans les bases de données de séquences. Bien que de nombreuses précautions seront à prendre en raison des problèmes engendrés par les annotations automatiques ou les conditions d'échantillonnage, ces bases de données sont un formidable vivier d'informations, sous-exploité, mais auquel on pourrait, en effet, s'intéresser dans l'avenir.

Quoi qu'il en soit, et ce n'est pas pour me déplaire, toutes ces perspectives pourraient offrir encore de belles années à venir à la multi-disciplinarité, à la bio-informatique et à la modélisation...

RÉFÉRENCES BIBLIOGRAPHIQUES

- Anderson RM, May RM. 1991. *Infectious Diseases of Humans*. Oxford Science Publications, Great Britain.
- Baranton G, Martino S De. 2009. *Borrelia burgdorferi* sensu lato diversity and its influence on pathogenicity in humans. *Curr. Probl. Dermatol.* 37:1–17.
- Baranton G, Postic D, Saint-Girons I, Boerlin P, Piffaretti J, Assous M, Grimont P. 1992. Delineation of *Borrelia burgdorferi* sensu stricto, *Borrelia garinii* sp. nov., and group VS461 associated with Lyme borreliosis. *Int. J. Syst. Bacteriol.* 42:378–383.
- Barbour A, Travinsky B. 2010. Evolution and distribution of the *ospC* gene, a transferable serotype determinant of *Borrelia burgdorferi*. *MBio* 1:1–11.
- Beaumont M. 2005. Adaptation and speciation: what can F_{st} tell us? *Trends Ecol. Evol.* 20:435–440.
- Beaumont M, Rannala B. 2004. The Bayesian revolution in genetics. *Nat. Rev. Genet.* 5:251–261.
- Beaumont M. 2010. Approximate Bayesian Computation in Evolution and Ecology. *Annu. Rev. Ecol. Evol. Syst.* 41:379–406.
- Begon M. 2008. Effects of host diversity on disease dynamics, p. 12–19. In Ostfeld, RS, Keesing, F, Eviner, VT (eds.), *Infectious Disease Ecology*. Princeton University Press, Princeton.
- Begon M, Bennett M, Bowers RG, French NP, Hazel SM, Turner J. 2002. A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiol. Infect.* 129:147–153.
- Bellet-Edimo R, Betschart B, Gern L. 2005. Frequency and efficiency of transovarial and subsequent transstadial transmissions of *Borrelia burgdorferi* sensu lato in *Ixodes ricinus* Ticks. *Bull. la Société Neuchâteloise des Sci. Nat.* 128:117–125.
- Benach J, Bosler E, Hanrahan JP, Coleman JL, Habicht GS, Bast TF, Cameron DJ, Ziegler JL, Barbour AG, Burgdorfer W, R Edelman RA. 1983. Spirochetes isolated from the blood of two patients with Lyme disease. *N. Engl. J. Med.*
- Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* 19:2609–2625.
- Biek R. 2007. Evolutionary dynamics and spatial genetic structure of epizootic hemorrhagic disease virus in the eastern United States. *Infect. Genet. Evol.* 7:651–655.
- Boers SA, van der Reijden WA, Jansen R. 2012. High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS ONE* 7(7): e39630.

Références bibliographiques

- doi:10.1371/journal.pone.0039630.
- Brisson D, Drecktrah D, Eggers CH, Samuels DS. 2012. Genetics of *Borrelia burgdorferi*. Annu. Rev. Genet. 46:515–36.
- Brisson D, Dykhuizen DE. 2004. *ospC* diversity in *Borrelia burgdorferi*: different hosts are different niches. Genetics 168:713–722.
- Brisson D, Dykhuizen DE. 2006. A modest model explains the distribution and abundance of *Borrelia burgdorferi* strains. Am. J. Trop. Med. Hyg. 74:615–622.
- Brisson D, Dykhuizen DE, Ostfeld RS. 2008. Conspicuous impacts of inconspicuous hosts on the Lyme disease epidemic. Proc. R. Soc. B Biol. Sci. 275:227–235.
- Bryce J, Boschi-Pinto C, Shibuya K, Black R. 2005. WHO estimates of the causes of death in children. Lancet 365:1147–1152.
- Burgdorfer W, Barbour A, Hayes S. 1983. Erythema chronicum migrans- A tickborne spirochetosis. Acta Trop. 69–77.
- Burgdorfer W, Barbour A, Hayes S, Benach J, Grunwaldt E, Davis J. 1982. Lyme disease-a tick-borne spirochetosis? Science. 216:1317–1319.
- Cadenas F, Rais O, Humair P, Douet V, Moret J, Gern L. 2007. Identification of host bloodmeal source and *Borrelia burgdorferi* sensu lato in field-collected *Ixodes ricinus* ticks in Chaumont (Switzerland). J. Med. Entomol. 44:1109–1117.
- Calow P. 1998. The Encyclopedia of Ecology and Environmental Management.
- Canica M, Nato F, Merle L, Mazie J, Baranton G, Postic D. 1993. Monoclonal antibodies for identification of *Borrelia afzelii* sp. nov. associated with late cutaneous manifestations of Lyme borreliosis. Scand. J. Infect. Dis. 25:441–448.
- Casjens SR, Fraser-Liggett CM, Mongodin EF, Qiu W-G, Dunn JJ, Luft BJ, Schutzer SE. 2011a. Whole genome sequence of an unusual *Borrelia burgdorferi* sensu lato isolate. J. Bacteriol. 193:1489–90.
- Casjens SR, Mongodin EF, Qiu W-G, Dunn JJ, Luft BJ, Fraser-Liggett CM, Schutzer SE. 2011b. Whole-genome sequences of two *Borrelia afzelii* and two *Borrelia garinii* Lyme disease agent isolates. J. Bacteriol. 193:6995–6996.
- Chapuis J, Obolenskaya E, Pisanu B, Lissovsky A. 2011. Datasheet on *Tamias sibiricus*. CABI, Wallingford, UK.
- Clark K, Hendricks A, Burge D. 2005. Molecular Identification and Analysis of *Borrelia burgdorferi* Sensu Lato in Lizards in the Southeastern United States. Appl. Environ. Microbiol. 71:2616–2625.

- Cleaveland S, Laurenson MK, Taylor LH. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 356:991–9.
- Comstedt P, Bergström S, Olsen B, Garpmo U, Marjavaara L, Mejlon H, Barbour AG, Bunikis J. 2006. Migratory passerine birds as reservoirs of Lyme borreliosis in Europe. *Emerg. Infect. Dis.* 12:1087–1095.
- Cooper TF, Heinemann J a. 2005. Selection for plasmid post-segregational killing depends on multiple infection: evidence for the selection of more virulent parasites through parasite-level competition. *Proc. R. Soc. B Biol. Sci.* 272:403–10.
- Crowder CD, Matthews HE, Schutzer S, Rounds M a, Luft BJ, Nolte O, Campbell SR, Phillipson C a, Li F, Sampath R, Ecker DJ, Eshoo MW. 2010. Genotypic variation and mixtures of Lyme *Borrelia* in *Ixodes* ticks from North America and Europe. *PLoS ONE* 5(5): e10650. doi:10.1371/journal.pone.0010650.
- de Roode JC, Pansini R, Cheesman SJ, Helinski MEH, Huijben S, Wargo AR, Bell AS, Chan BHK, Walliker D, Read AF. 2005. Virulence and competitive ability in genetically diverse malaria infections. *Proc. Natl. Acad. Sci. U. S. A.* 102:7624–7628.
- Diallo M, Thonnon J, Traore-Lamizana M, Fontenille D. 1999. Vectors of Chikungunya virus in Senegal: current data and transmission cycles. *Am. J. Trop. Med. Hyg.* 60:281–286.
- Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.
- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.
- Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7.
- Dubska L, Literak I, Kocianova E, Taragelova V, Sychra O. 2009. Differential role of passerine birds in distribution of *Borrelia* spirochetes, based on data from ticks collected from birds during the postbreeding migration period in Central Europe. *Appl. Environ. Microbiol.* 75:596–602.
- Dupuy V, Manso-Silván L, Barbe V, Thebault P, Dordet-Frisoni E, Citti C, Poumarat F, Blanchard A, Breton M, Sirand-Pugnet P, Thiaucourt F. 2012. Evolutionary history of contagious bovine pleuropneumonia using next generation sequencing of *Mycoplasma mycoides* Subsp. *mycoides* “Small Colony”. *PLoS ONE* 7(10): e46821.

Références bibliographiques

- doi:10.1371/journal.pone.0046821.
- Dykhuizen DE, Baranton G. 2001. The implications of a low rate of horizontal transfer in *Borrelia*. Trends Microbiol. 9:344–350.
- Falco R, Daniels T, Fish D. 1995. Increase in abundance of immature *Ixodes scapularis* (Acari: Ixodidae) in an emergent Lyme disease endemic area. J. Med. Entomol. 32:522–526.
- Fay J, Wu C. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. 16:1003–1005.
- Forrester NL, Guerbois M, Seymour RL, Spratt H, Weaver SC. 2012. Vector-borne transmission imposes a severe bottleneck on an RNA virus population. PLoS Pathog 8(9): e1002897. doi:10.1371/journal.ppat.1002897.
- Foxman B, Riley L. 2001. Molecular epidemiology: focus on infection. Am. J. Epidemiol. 153:1135–1141.
- Frank S. 1992. A kin selection model for the evolution of virulence. Proc. R. Soc. B Biol. Sci. 250:195–197.
- Frank S. 1996. Models of Parasite Virulence. Q. Rev. Biol. 71:37–78.
- Franke J, Hildebrandt A, Dorn W. 2013. Exploring gaps in our knowledge on Lyme borreliosis spirochaetes--updates on complex heterogeneity, ecology, and pathogenicity. Ticks Tick. Borne. Dis. 4:11–25.
- Fraser C, Casjens S, Huang W, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb J-F, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, Vugt R van, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Wathley L, McDonald L, Artiach P, Bowman C, Garland S, Fujii C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580–588.
- Frischknecht F, Baldacci P, Martin B, Zimmer C, Thiberge S, Olivo-Marin J-C, Shorte SL, Ménard R. 2004. Imaging movement of malaria parasites during transmission by Anopheles mosquitoes. Cell. Microbiol. 6:687–694.
- Gago S, Elena SF, Flores R, Sanjuán R. 2009. Extremely high mutation rate of a hammerhead viroid. Science. 323:1308.
- Gern L, Rais O. 1996. Efficient transmission of *Borrelia burgdorferi* between cofeeding

- Ixodes ricinus* ticks (Acari: Ixodidae). J. Med. Entomol. 33:189–192.
- Girardin P, Brossard M. 1990. Rabbits infested with *Ixodes ricinus* adults: effects of a treatment with cyclosporin A on the biology of ticks fed on naive and immune hosts. Ann. Parasitol. Hum. Comparée 65:262–266.
- Glöckner G, Lehmann R, Romualdi A, Pradella S, Schulte-Spechtel U, Schilhabel M, Wilske B, Sühnel J, Platzer M. 2004. Comparative analysis of the *Borrelia garinii* genome. Nucleic Acids Res. 32:6038–6046.
- Grünwald NJ, Goss EM. 2011. Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches. Annu. Rev. Phytopathol. 49:5.1–5.19.
- Guttman D, Wang P, Wang I-N, Bosler EMB, Luft BJ, Dykhuizen DE. 1996. Multiple infections of *Ixodes scapularis* ticks by *Borrelia burgdorferi* as revealed by single-strand conformation polymorphism analysis. J. Clin. Microbiol. 34:652–656.
- Hanincová K, Taragelová V, Koci J, Schäfer SM, Hails R, Ullmann AJ, Piesman J, Labuda M, Kurtenbach K. 2003. Association of *Borrelia garinii* and *B. valaisiana* with songbirds in Slovakia. Appl. Environ. Microbiol. 69:2825–2830.
- Harrison F, Browning LE, Vos M, Buckling A. 2006. Cooperation and virulence in acute *Pseudomonas aeruginosa* infections. BMC Biol. 4.
- Haven J, Vargas LC, Mongodin EF, Xue V, Hernandez Y, Pagan P, Fraser-Liggett CM, Schutzer SE, Luft BJ, Casjens SR, Qiu W-G. 2011. Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in *Borrelia burgdorferi*, the lyme disease bacterium. Genetics 189:951–966.
- Haydon DT, Cleaveland S, Taylor LH, Laurenson MK. 2002. Identifying reservoirs of infection: a conceptual and practical challenge. Emerg. Infect. Dis. 8:1468–1473.
- Hayes EB, Komar N, Nasci RS, Montgomery SP, O’Leary DR, Campbell GL. 2005. Epidemiology and transmission dynamics of West Nile virus disease. Emerg. Infect. Dis. 11:1167–1173.
- Hinnebusch J, Barbour A. 1992. Linear-and circular-plasmid copy numbers in *Borrelia burgdorferi*. J. Bacteriol. 174:5251–5257.
- Hu CM, Humair P-F, Wallich R, Gern L. 1997. Apodemus sp. rodents, reservoir hosts for *Borrelia afzelii* in an endemic area in Switzerland. Zentralblatt für Bakteriologie. 285:558–564.
- Hudson RR. 1991. Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol.

Références bibliographiques

- 7:1–44.
- Hudson RR. 1994. Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *J. Evol. Biol.* 7:535–548.
- Humair PF, Rais O, Gern L. 1999. Transmission of *Borrelia afzelii* from *Apodemus* mice and *Clethrionomys* voles to *Ixodes ricinus* ticks: differential transmission pattern and overwintering maintenance. *Parasitology* 118:33–42.
- Humair P, Douet V, Cadenas F, Schouis L, Van De Pol I, Gern L. 2007. Molecular identification of bloodmeal source in *Ixodes ricinus* ticks using 12S rDNA as a genetic marker. *J. Med. Entomol.* 44:869–880.
- Humair P, Gern L. 1998. Relationship between *Borrelia burgdorferi* sensu lato species, red squirrels (*Sciurus vulgaris*) and *Ixodes ricinus* in enzootic areas in Switzerland. *Acta Trop.* 69:213–227.
- Humair P, Postic D, Wallich R, Gern L. 1998. An avian reservoir (*Turdus merula*) of the Lyme borreliosis spirochetes. *Zentralblatt für Bakteriologie* 538:521–538.
- Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254–267.
- Hyde F, Johnson R. 1984. Genetic relationship of lyme disease spirochetes to *Borrelia*, *Treponema*, and *Leptospira* spp. *J. Clin. Microbiol.* 20:151–154.
- Johnson PTJ, Thielges DW. 2010. Diversity, decoys and the dilution effect: how ecological communities affect disease risk. *J. Exp. Biol.* 213:961–70.
- Johnson PTJ, Hartson RB, Larson DJ, Sutherland DR. 2008. Diversity and disease: community structure drives parasite transmission and host fitness. *Ecol. Lett.* 11:1017–26.
- Jones KE, Patel NG, Levy M, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008. Global trends in emerging infectious diseases. *Nature* 451:990–993.
- Keesing F, Belden LK, Daszak P, Dobson A, Harvell CD, Holt RD, Hudson P, Jolles A, Jones KE, Mitchell CE, Myers SS, Bogich T, Ostfeld RS. 2010. Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* 468:647–52.
- Kempf F, De Meeûs T, Vaumourin E, Noel V, Taragel'ová V, Plantard O, Heylen DJ, Eraud C, Chevillon C, McCoy KD. 2011. Host races in *Ixodes ricinus*, the European vector of Lyme borreliosis. *Infect. Genet. Evol.* 11:2043–2048.
- Kent RJ. 2009. Molecular methods for arthropod bloodmeal identification and applications to ecological and vector-borne disease studies. *Mol. Ecol. Resour.* 9:4–18.

- Kurtenbach K, Dizij A, Seitz H, Margos G, Moter S, Kramer D, Wallich R, Shaible U, Simon M. 1994. Differential immune responses to *Borrelia burgdorferi* in European wild rodent species influence spirochete transmission to *Ixodes ricinus* L.(Acari: Ixodidae). *Infect. Immun.* 62:5344–5352.
- Kurtenbach K, Peacey M, Rijpkema SG, Hoodless AN, Nuttall PA, Randolph SE. 1998. Differential transmission of the genospecies of *Borrelia burgdorferi* sensu lato by game birds and small rodents in England. *Appl. Environ. Microbiol.* 64:1169–1174.
- Kurtenbach K, Hanincová K, Tsao JI, Margos G, Fish D, Ogden NH. 2006. Fundamental processes in the evolutionary ecology of Lyme borreliosis. *Nat. Rev. Microbiol.* 4:660–9.
- Kurtenbach K, Michelis S, Etti S, Schäfer SM, Sewell H-S, Brade V, Kraiczy P. 2002. Host association of *Borrelia burgdorferi* sensu lato—the key role of host complement. *Trends Microbiol.* 10:74–79.
- Lee K, Klasing KC. 2004. A role for immunology in invasion biology. *Trends Ecol. Evol.* 19:523–529.
- Lewontin R. 1964. The interaction of selection and linkage. *Genetics* 50:757–782.
- Lipschütz B. 1913. Über eine seltene Erythemform (*Erythema chronicum migrans*). *Arch. Dermatol. Res.*
- Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* 95:3140–5.
- Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* 11:728–736.
- Margos G, Vollmer S, Cornet M, Garnier M, Fingerle V, Wilske B, Bormane A, Vitorino L, Collares-Pereira M, Drancourt M, Kurtenbach K. 2009. A new *Borrelia* species defined by multilocus sequence analysis of housekeeping genes. *Appl. Environ. Microbiol.* 75:5410–5416.
- Margos G, Vollmer S, Ogden NH, Fish D. 2011. Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. *Infect. Genet. Evol.* 11:1545–1563.
- Marsot M, Sigaud M, Chapuis JL, Ferquel E, Cornet M, Vourc'h G. 2011. Introduced Siberian chipmunks (*Tamias sibiricus barberi*) harbor more-diverse *Borrelia*

Références bibliographiques

- burgdorferi* sensu lato genospecies than native bank voles (*Myodes glareolus*). Appl. Environ. Microbiol. 77:5716–521.
- Marsot M. 2011. Modification du risque d'une maladie multi-hôtes suite à l'introduction d'une espèce réservoir : Cas de la maladie de Lyme et du tamia de Sibérie en Ile-de-France.
- Marsot M, Chapuis J-L, Gasqui P, Dozières A, Masségli S, Pisanu B, Ferquel E, Vourc'h G. 2013. Introduced Siberian chipmunks (*Tamias sibiricus barberi*) contribute more to lyme borreliosis risk than native reservoir rodents. PLoS ONE 8(1): e55377. doi:10.1371/journal.pone.0055377.
- Matic I, Rayssiguier C, Radman M. 1995. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell 80:507–515.
- Maynard Smith J, Feil EJ, Smith NH. 2000. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 22:1115–22.
- Maynard Smith J, Smith NH. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. 15:590–599.
- Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. U. S. A. 90:4384–4388.
- Mayxay M, Pukrittayakamee S, Newton PN, White NJ. 2004. Mixed-species malaria infections in humans. Trends Parasitol. 20:233–40.
- McCallum H, Barlow N, Hone J. 2001. How should pathogen transmission be modelled? Trends Ecol. Evol. 16:295–300.
- McCoy KD, Léger E, Dietrich M. 2013. Host specialization in ticks and transmission of tick-borne diseases: a review. Front. Cell. Infect. Microbiol. 3.
- Messina JP, Brady OJ, Scott TW, Zou C, Pigott DM, Duda KA, Bhatt S, Katzelnick L, Howes RE, Battle KE, Simmons CP, Hay SI. 2014. Global spread of dengue virus types: mapping the 70 year history. Trends Microbiol. 22:138–46.
- Mongodin EF, Casjens SR, Bruno JF, Xu Y, Drabek EF, Riley DR, Cantarel BL, Pagan PE, Hernandez YA, Vargas LC, Dunn JJ, Schutzer SE, Fraser CM, Qiu W-G, Luft BJ. 2013. Inter- and intra-specific pan-genomes of *Borrelia burgdorferi* sensu lato: genome stability and adaptive radiation. BMC Genomics 14:1–25.
- Morlon H, Kems BD, Plotkin JB, Brisson D. 2012. Explosive radiation of a bacterial species group. Evolution 66:2577–2586.
- Nordborg M. 2000. Coalescent Theory: An Introduction. Syst. Biol. 58:162–165.

- Nowak M, May R. 1994. Superinfection and the evolution of parasite virulence. *Proc. R. Soc. B Biol. Sci.* 255:81–89.
- Olsén B, Jaenson T, Bergström S. 1995. Prevalence of *Borrelia burgdorferi* sensu lato-infected ticks on migrating birds. *Appl. Environ. Microbiol.* 61:3082–3087.
- Ostfeld RS. 2009. Biodiversity loss and the rise of zoonotic pathogens. *Clin. Microbiol. Infect. Eur. Soc. Clin. Microbiol. Infect. Dis.* 15:40–43.
- Ostfeld RS, Keesing F. 2000. The function of biodiversity in the ecology of vector-borne zoonotic diseases. *Can. J. Zool.* 78:2061–2078.
- Pérez D, Kneubühler Y, Rais O, Gern L. 2012. Seasonality of *Ixodes ricinus* ticks on vegetation and on rodents and *Borrelia burgdorferi* sensu lato genospecies diversity in two lyme borreliosis-endemic areas in Switzerland. *Vector borne zoonotic Dis.* 12:633–44.
- Piesman J, Gern L. 2004. Lyme borreliosis in Europe and North America. *Parasitology* 129:S191–S220.
- Pisanu B, Marsot M, Marmet J, Chapuis J-L, Réale D, Vourc'h G. 2010. Introduced Siberian chipmunks are more heavily infested by ixodid ticks than are native bank voles in a suburban forest in France. *Int. J. Parasitol.* 40:1277–1283.
- Postic D, Edlinger C, Richaud C, Grimont F, Dufresne Y, Perolat P, Baranton G, Grimont P. 1990. Two genomic species in *Borrelia burgdorferi*. *Res. Microbiol.* 141:465–475.
- Poulin R, Combes C. 1999. The concept of virulence: interpretations and implications. *Parasitol. Today* 15:474–475.
- Qiu W-G, Bosler EM, Campbell JR, Ugine GD, Wang I-N, Luft BJ, Dykhuizen DE. 1997. A population genetic study of *Borrelia burgdorferi* sensu stricto from eastern Long Island, New York, suggested frequency-dependent selection, gene flow and host adaptation. *Hereditas* 127:203–216.
- Qiu W-G, Schutzer SE, Bruno JF, Attie O, Xu Y, Dunn JJ, Fraser CM, Casjens SR, Luft BJ. 2004. Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing. *Proc. Natl. Acad. Sci. U. S. A.* 101:14150–14155.
- Qiu W, Dykhuizen D, Acosta M, Luft B. 2002. Geographic uniformity of the Lyme disease spirochete (*Borrelia burgdorferi*) and its shared history with tick vector (*Ixodes scapularis*) in the northeastern United States. *Genetics* 160:833–849.
- Rafalski A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr.*

Références bibliographiques

- Opin. Plant Biol. 5:94–100.
- Randolph SE, Craine NG. 1995. General framework for comparative quantitative studies on transmission of tick-borne diseases using Lyme borreliosis in Europe as an example. *J. Med. Entomol.* 32:765–777.
- Rappole J, Compton BW, Leimgruber P, Robertson J, King DI, Renner SC. 2006. Modeling movement of West Nile virus in the Western hemisphere. *Vector-Borne Zoonotic Dis.* 6:128–139.
- Reluga T, Meza R, Walton D, Galvani A. 2007. Reservoir interactions and disease emergence. *Theor. Popul. Biol.* 72:400–408.
- Richter D, Schlee D, Allgöwer R, Matuschka F. 2004. Relationships of a novel Lyme disease spirochete, *Borrelia spielmani* sp. nov., with its hosts in Central Europe. *Appl. Environ. Microbiol.* 70:6414–6419.
- Schmidt K, Ostfeld R. 2001. Biodiversity and the dilution effect in disease ecology. *Ecology* 82:609–619.
- Schorderet S, Brossard M. 1993. Changes in immunity to *Ixodes ricinus* by rabbits infested at different levels. *Med. Vet. Entomol.*
- Schutzer SE, Fraser-Liggett CM, Casjens SR, Qiu W-G, Dunn JJ, Mongodin EF, Luft BJ. 2011. Whole-genome sequences of thirteen isolates of *Borrelia burgdorferi*. *J. Bacteriol.* 193:1018–1020.
- Schutzer SE, Fraser-Liggett CM, Qiu W-G, Kraiczy P, Mongodin EF, Dunn JJ, Luft BJ, Casjens SR. 2012. Whole-genome sequences of *Borrelia bissettii*, *Borrelia valaisiana*, and *Borrelia spielmanii*. *J. Bacteriol.* 194:545–546.
- Skuballa J, Oehme R, Hartelt K, Petney T, Bücher T, Kimmig P, Taraschewski H. 2007. European Hedgehogs as Hosts for *Borrelia* spp., Germany. *Emerg. Infect. Dis.* 13:952–953.
- Slatkin M. 1985. Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* 16:393–430.
- Spratt BG, Hanage WP, Feil EJ. 2001. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr. Opin. Microbiol.* 4:602–606.
- Stanek G, Wormser GP, Gray J, Strle F. 2012. Lyme borreliosis. *Lancet* 379:461–73.
- Stapleford KA, Coffey LL, Lay S, Bordería AV, Duong V, Isakov O, Rozen-Gagnon K, Arias-Goeta C, Blanc H, Beaucourt S, Haliloğlu T, Schmitt C, Bonne I, Ben-Tal N, Shomron N, Failloux A-B, Buchy P, Vignuzzi M. 2014. Emergence and transmission of arbovirus evolutionary intermediates with epidemic potential. *Cell Host Microbe*

- 15:706–716.
- Steere A, Malawista S, Snyderman D, Shope R, Andiman W, Ross M, Steel F. 1977. Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three connecticut communities. *Arthritis Rheum* 20:7–17.
- Steere A, Coburn J, Glickstein L. 2004. The emergence of Lyme disease. *J. Clin. Invest.* 113:1093–1101.
- Stricker RB, Johnson L. 2014. Lyme disease: call for a “Manhattan Project” to combat the epidemic. *PLoS Pathog* 10(1): e1003796. doi:10.1371/journal.ppat.1003796.
- Sutherst R. 2004. Global change and human vulnerability to vector-borne diseases. *Clin. Microbiol. Rev.* 17:136–173.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–95.
- Tälleklint L, Jaenson TGT. 1995. Is the Small Mammal (*Clethrionomys glareolus*) or the Vector (*Ixodes ricinus*) the Primary Overwintering Reservoir for the Lyme Borreliosis Spirochete in Sweden? *J. Wildl. Dis.* 31:537–540.
- Tälleklint L, Jaenson T. 1994. Transmission of *Borrelia burgdorferi* s.l. from mammal reservoirs to the primary vector of Lyme borreliosis, *Ixodes ricinus* (Acari: Ixodidae), in Sweden. *J. Med. Entomol.* 31:880–886.
- Telford 3rd S, Mather T, Moore S, Wilson M, Spielman A. 1988. Incompetence of deer as reservoirs of the Lyme disease spirochete. *Am. J. Trop. Med. Hyg.* 39:105–109.
- Timms R, Read A. 1999. What makes a specialist special? *Trends Ecol. Evol.* 14:333–334.
- Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf M. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6:187–202.
- Tsao K, Bent SJ, Fish D. 2013. Identification of *Borrelia burgdorferi ospC* genotypes in host tissue and feeding ticks by terminal restriction fragment length polymorphisms. *Appl. Environ. Microbiol.* 79:958–964.
- Velthuis A, Saatkamp H, Mourits M, de Koeijer A, Elbers A. 2010. Financial consequences of the Dutch bluetongue serotype 8 epidemics of 2006 and 2007. *Prev. Vet. Med.* 93:294–304.
- Viana M, Mancy R, Biek R, Cleaveland S, Cross PC, Lloyd-Smith JO, Haydon DT. 2014. Assembling evidence for identifying reservoirs of infection. *Trends Ecol. Evol.* 29:270–279.

Références bibliographiques

- Vollmer SA, Bormane A, Dinnis RE, Seelig F, Dobson ADM, Aanensen DM, James MC, Donaghy M, Randolph SE, Feil EJ, Kurtenbach K, Margos G. 2011. Host migration impacts on the phylogeography of Lyme Borreliosis spirochaete species in Europe. *Environ. Microbiol.* 13:184–192.
- Volz EM, Koelle K, Bedford T. 2013. Viral phylodynamics. *PLoS Comput Biol* 9(3): e1002947. doi:10.1371/journal.pcbi.1002947.
- Wang G, Dam A Van, Le Fleche A, Postic D, Peter O, Baranton G, De Boer R, Spanjaard L, Dankert J. 1997. Genetic and phenotypic analysis of *Borrelia valaisiana* sp. nov. (*Borrelia* genomic groups VS116 and M19). *Int. J. Syst. Bacteriol.* 47:926–932.
- Whitlock M. 1996. The red queen beats the jack-of-all-trades: the limitations on the evolution of phenotypic plasticity and niche breadth. *Am. Nat.* 148:S65–S77.
- WHO. 2009. Dengue, Guidelines for diagnosis, treatment, prevention and control.
- Wiuf C, Hein J. 2000. The coalescent with gene conversion. *Genetics* 155:451–462.
- Woolhouse MEJ, Dye C. 2001. Population biology of emerging and re-emerging pathogens. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 356:981–1106.
- Woolhouse MEJ. 2002. Population biology of emerging and re-emerging pathogens. *Trends Microbiol.* 10:S3–7.
- Woolhouse MEJ, Taylor LH, Haydon DT. 2001. Population Biology of Multihost Pathogens. *Science.* 292:1109–1112.
- Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. sixth Int. Congr. Genet.* 356–366.

Diversité génomique des bactéries pathogènes du complexe d'espèces *Borrelia burgdorferi* : évolution et épidémiologie moléculaire

Les maladies infectieuses sont une des causes les plus importantes de morbidité chez l'homme et l'animal avec des conséquences à la fois économiques, sanitaires et écologiques. L'étude de la diversité des pathogènes responsables et de leurs dynamiques de circulation au sein des communautés d'hôtes et de vecteurs, peut fournir des informations importantes pour la prévention et le contrôle de ces maladies. Dans ce contexte, nous nous sommes intéressés à l'agent pathogène responsable de la maladie de Lyme. Cette maladie est causée par les bactéries du complexe d'espèces *Borrelia burgdorferi* sensu lato (s.l.) transmises par les tiques lors de repas sanguins et sont capables d'infecter plusieurs espèces d'hôtes vertébrés. L'analyse de la diversité génétique de 63 souches de *B. burgdorferi* s.l., dont les génomes ont été séquencés, ont révélé un degré d'isolement génétique très important entre les différentes espèces du complexe. Les résultats obtenus suggèrent que les différents spectres d'hôtes des lignées de *B. burgdorferi* s.s. (principalement associées aux petits mammifères) et de *B. garinii* (normalement associées aux oiseaux) conduisent à des dynamiques de populations distinctes. De plus, grâce au séquençage haut-débit de deux marqueurs, nous avons pu démontrer qu'il existe, à une échelle intra-spécifique, des associations préférentielles des génotypes de *B. burgdorferi* avec différentes espèces de rongeurs. Enfin, en utilisant la diversité observée chez ces rongeurs et celle chez les tiques, nous avons estimé, *via* une approche de modélisation, que la contribution au risque de la maladie pour l'homme d'une espèce hôte introduite (tamia de Sibérie), pouvait être importante.

Mots-clés : *B. burgdorferi* s.l., tiques, hôtes, diversité génétique, dynamique de populations, épidémiologie, spécificité d'hôte

Genomic diversity of pathogenic bacteria in the *Borrelia burgdorferi* species complex: evolution and molecular epidemiology

Infectious diseases are one of the major causes of human and animal morbidity, and they have impacts on the economy, public health, and the environment. By studying the diversity of the pathogens responsible for these diseases and their circulation within host communities and among vectors, we may glean valuable information that will aid prevention and control efforts. For these reasons, during my thesis, I became particularly interested in the pathogen(s) responsible for Lyme disease. This disease is caused by bacteria belonging to the *Borrelia burgdorferi* sensu lato (s.l.) species complex that are transmitted by ticks (during their blood meals) and that can infect several vertebrate host species. When I analyzed the genetic diversity present in 63 *B. burgdorferi* s.l. strains, whose genomes had been sequenced, I found that there was a significant degree of genetic separation among the different genospecies making up the complex. My results suggest that the fact that these different bacterial groups infect different ranges of hosts—*B. burgdorferi* s.s. is mainly a pathogen of small mammals and *B. garinii* is primarily associated with birds—lead to distinct population dynamics. Moreover, thanks to the high-throughput sequencing of two genetic markers, I have been able to show that, at an intraspecific level, certain *B. burgdorferi* genotypes are associated with specific rodent species. Finally, using the pathogen diversity observed in rodents and ticks, I employed a modeling approach to estimate the human disease risks presented by an introduced host species (the Siberian chipmunk) and found that these risks could be significant.

Key words: *B. burgdorferi* s.l., ticks, hosts, genetic diversity, population dynamics, epidemiology, host specificity

Discipline : Physiologie et génétique moléculaires (Biologie évolutive et épidémiologie)

Laboratoire :

Unité d'Épidémiologie Animale-UR346
Institut National de la Recherche Agronomique
Centre de Recherche de Clermont-Ferrand – Theix
63122 Saint Genès-Champanelle